The EU Commission’s proposed TDM exception: unlocking research or innovation?

Candidate number: 7010
Submission deadline: 30.06.17
Number of words: 16380
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.1</td>
<td>Big Data and research practices</td>
<td>3</td>
</tr>
<tr>
<td>1.2</td>
<td>Methods</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>On the history of machine reading</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>Defining the phenomenon</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>Technological models of TDM</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Definitions of the term TDM in the literature</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>On the current exceptions</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Copyright law</td>
<td>20</td>
</tr>
<tr>
<td>3.2</td>
<td>Database law</td>
<td>25</td>
</tr>
<tr>
<td>4</td>
<td>Assessing the adequacy</td>
<td>27</td>
</tr>
<tr>
<td>4.1</td>
<td>Japan</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>The UK</td>
<td>29</td>
</tr>
<tr>
<td>4.3</td>
<td>The EU</td>
<td>34</td>
</tr>
<tr>
<td>5</td>
<td>Conclusions</td>
<td>39</td>
</tr>
<tr>
<td>6</td>
<td>References</td>
<td>40</td>
</tr>
</tbody>
</table>
1 Introduction
1.1 Big Data and research practices

The term Big Data in its current use was coined by Roger Magoulas in 2005 and refers to a wide range of large data sets that due to their size and complexity are almost impossible to manage and process using traditional data management tools.¹ The key enabling factor in the growth of digital data is the decrease in data storage cost. Other enablers of Big Data are the spread of mobile devices, sensors and sensor networks, and the Internet of Things (IoT).² From the technological point of view, IoT can be viewed as billions of devices and components with an internet address, enabling them to communicate in massive sensing systems.³ The phenomenon of Big Data spans three different dimensions – volume, velocity and variety, meaning that huge amounts of data is generated at a very fast pace from different sources.⁴ According to IBM, every day we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created during the last two years.⁵ Facebook users alone share over 30 million pieces of content per month, and Twitter has 350 million tweets per day. Almost all scientific journals are already available online and 2.5 million scientific articles are published every year.⁶ Having so many articles available has led to the problem that our traditional ability to analyze data has been far outstripped by the shear amount of data available.

Scientists had to come up with alternative research practices to overcome challenges posed by Big Data. The research practice that is colloquially known as ‘Text and Data Mining’ (TDM) covers a range of techniques that allows the researcher to analyze vast amounts of data to discover new knowledge. TDM is not a new technology, but it has received spotlight attention due to proliferation of Big Data. The applications of TDM are very diverse and span multiple disciplines. Healthcare and medical research are the preeminent areas to benefit from TDM activities. The technology allows scientists to create new copyrighted works that contain new information, which is based on data that are already available, but which could not previously be inspected so rigorously. Another area is computer science itself, where the length of code and its complexity have led to a situation where complex systems are difficult to debug⁷.⁸ The McKinsey study

---

¹ Halevi and Moed (2012)
² Organisation for Economic Co-operation and Development (2013) p. 8
³ Hargreaves (2011) p. 14
⁴ Gartner, Inc. (2011)
⁵ IBM (2014)
⁶ Ware and Mabe (2015) p. 6
⁷ Debugging is the process of finding and resolving of defects in software.
⁸ Thorburn et al. (2017) p.131
indicates that TDM could help European economies to save more than 100 billion euro in operational efficiency by using data more effectively.\(^9\)

Despite the benefits that TDM can provide for the research community and the EU economy, the report of Sergey Filippov shows that EU researchers use TDM less frequently than their counterparts from Asia and the US.\(^10\) The report from the expert group\(^11\) chaired by Ian Hargreaves outlines many factors that impede TDM activities in Europe, \textit{inter alia}, the lack of clarity around the legality of TDM.\(^12\) The legal ambiguity stems from the fact that TDM involves copying of the whole or a substantial part of work or database. Original works are protected in the EU under 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 (the InfoSoc Directive). Generally, the EU copyright law does not allow an unauthorized copying of a copyrighted work. In the case brought before the ECJ by Infopaq International A/S against Danske Dagblades Forening, the European Court of Justice (the ECJ) held that copying of even an eleven-word snippet could infringe copyright.\(^13\) Infopaq case illustrates how even a minor act of reproduction can result in a copyright infringement. Databases that are used by researchers for TDM purposes are protected under Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (the Database Directive). If the content of a database is qualified as work, such database is granted protection under both copyright and database law.

Fortunately, low thresholds for copyright or database \textit{sui generis} infringement that are stipulated in the InfoSoc Directive and Database directive are mitigated by the fact that both directives provide exhaustive lists of exceptions that may render copying of an entire subject-matter as lawful. However, it is not evident from the wording of the relevant exceptions whether they are applicable in the case of TDM. In addition, it is not clear whether legal access to the digital materials is sufficient to ensure lawfulness of TDM activities that are carried out on these materials. A license agreement between a publisher and a user creates a legal bases for lawful access to copyrighted materials. However, many of these agreements do not state whether they permit TDM. The question arises – (i) if the user has already entered into a license agreement with a publisher and accesses the materials lawfully, can he or she carry out TDM activities on these materials? Other questions that the reader may wonder about after he or she has been briefly

---

\(^9\) Manyika et al. (2011) p.2
\(^10\) Filippov (2014) p. 23
\(^11\) The expert group was set up by the EU Commission to obtain an advice from outside experts as a basis for sound policymaking.
\(^12\) Hargreaves et al. (2014) p.7
\(^13\) Infopaq International A/S v. Danske Dagblades Forening (2009), par. 66-70
introduced to the issue of TDM are: (ii) how does law regulate reprography of copyrighted materials carried out for purposes of a TDM-enabled research? (iii) can web data that is available to everyone be freely utilized for TDM?

In September 2016, the European Commission suggested a proposal for a directive, ‘On copyright in the Digital Single Market’ (the Digital Market Directive). Article 3 of the proposal provides for a TDM exception. In this paper, I will endeavor to answer the following research question –

Is Article 3 necessary, if so, does it address adequately the legal ambiguity that surrounds the technology?

Assessing adequacy, I will evaluate whether the exception:

(i) provides for an accurate and technology-neutral TDM definition;
(ii) covers all intellectual property rights (IPRs) that may be triggered by TDM processes;
(iii) includes all person and/or entities that should benefit from the exception.

1.2 Methods
Answering the research question, I will compare the EU proposal with TDM exemptions that exist in Japan and the UK. Since the research is concerned with comparing the legislation of different countries, adopting comparative law analysis becomes a sound methodological choice. This type of analysis might be viewed as a bundle of many mutually dependent methods, such as functional, structural and historical, to name just a few. Different ratios of these methods result in distinct instances of comparative law analysis, and therefore this kind of analysis varies greatly between research papers. I will try to maintain a balanced composition of my comparative law analysis so that none of the methods unduly prevails over the others. This approach should contribute positively to my research design. Regarding selection of legal systems to be compared, I will discuss jurisdictions that satisfy two criteria. Firstly, the legal system has to provide for an IPR exception that, which explicitly permits TDM activities. Secondly, the jurisdiction has to be prominent in the context of the international intellectual property regime. Concerning sources, I will examine constitutional and legislative rules, case law, doctrinal books and articles.

1.3 On the history of machine reading

The technological evolution brings challenges that are not always effectively addressed by the existing copyright legislation. There was substantial uncertainty as to the status of copies made for machines under the Berne Convention for the Protection of Literary and Artistic Works (the Berne Convention) and national legislation. Machine reading challenged copyright law for the first time in the late nineteenth century when the pianola\(^\text{15}\) first emerged. In 1899, the UK Court of Appeal in Boosey v. Whight\(^\text{16}\) considered whether perforated sheets that were used in the player piano were ‘copies’ of the plaintiff’s musical scores, and therefore infringing the copyright of the author of the music. The court decided that the perforated sheets were not ‘copies’ in a copyright sense, because they were made for a different purpose than to be read or played by persons. Commenting on this and on other cases brought before French and German courts, legal scholar William Briggs concluded that ‘the interests of the makers of mechanical musical instruments are not paramount, and should not be allowed to over-ride the ordinary right of composers’.\(^\text{17}\) The uncertainty as to the status of these machine-readable objects was solved with a legislative intervention. After the Berlin revision of 1908, the Berne Convention incorporated Article 13 that granted the authors the exclusive right to authorize adaptation of their musical works to instruments which can reproduce them mechanically, meaning that machine readable objects for the first time received the status of ‘copies’.

Article 13 of Berne Convention in the Berlin revision of 1908 reads as ‘[t]he authors of musical works shall have the exclusive right of authorizing: (1) the adaptation of those works to instruments which can reproduce them mechanically; (2) the public performance of the said works by means of these instruments […].’

Another time when machine reading captured the attention of the courts was with the introduction of personal computers. The question arose as to whether copyright subsists in the object code\(^\text{18}\) of a computer program. In the Australian case of Apple v. Edge, the court found that the object code is copyright protected as a ‘translation’ or as an ‘adaptation’ of the source code. However, Judge Sheppard J. dissented and held that copyright does not subsist in the object code because it is something that ‘only the microprocessor can “understand” or “see”’.\(^\text{19}\) The judge distinguished the case concerning the pianola on the basis that in that case the music could be heard when played on

\(^{15}\) Pianola is a player piano that can play automatically music that is recorded on perforated paper or metallic rolls.

\(^{16}\) Boosey v. Whight

\(^{17}\) Briggs (1906) p. 419

\(^{18}\) Object code is a sequence of instructions that usually consist of binary code that can be executed directly by the CPU. For humans object code appears as a long string of zeros and ones. In contrast, source code is a collection of computer instructions written using a human-readable programming language like Java or C++.

\(^{19}\) Apple Computer, Inc. v. Computer Edge Pty., Ltd.
a pianola, whereas object code is comprehensible only to a machine. Commenting on the Australian Copyright Act, Sheppard J. concluded that the subject of copyright has to be capable of being published and thus being seen or heard. The issue of object code was solved by adopting new legislation. In 1991, the European Council adapted Directive 91/250/EEC of 14 May 1991 on the legal protection of computer programs (the Computer Programs Directive). Article 1(2) of the directives reads,

“Protection in accordance with this Directive shall apply to the expression in any form of a computer program. Ideas and principles which underlie any element of a computer program, including those which underlie its interfaces, are not protected by copyright under this Directive”.

The wording “the expression in any form” allowed the EU lawmakers to bring into the scope of copyright protection an object code of a computer program. Later, wording similar to the Computer Programs Directive was included to other national and international laws.

The historical account above indicates that the legal ambiguity that hunted instances of machine reading that precede TDM came from the fact that the judges were reluctant to recognize copying for machines as a copyright infringing activity as copies that result from such copying were not meant for human apprehension and bore per se no expressive value.

2 Defining the phenomenon
2.1 Technological models of TDM

In this section, I will describe data analysis as a set of steps or a sum of technological processes to introduce the reader to the technological dimension of the phenomenon before taking on discussing definitions that endeavor to provide the holistic account of the activity. In addition, I will mention IPRs that are triggered when the researcher implements certain processes of data analysis. Finally,
I will touch upon ‘level of access’ because this notion is crucial for components such as contractual overridability and ‘lawful access’. These components are to be discussed in detail later in this paper as they are found in the enacted UK law and the EU draft legislation that regulates data analysis.

I will start with the digitalizing process since for the researcher to be able to analyze data contained in paper sources he or she has to digitalize those materials. Reprography is a technique to convert analog sources into digital materials that relies on creating a permanent reproduction of the work in a digital medium. Reprography is a form of reproduction, i.e. the duplication of a work. Reprographic reproduction is a process that usually results in a copy on a graphic surface by such process as printing, photocopying, scanning, digital copying (for instance on CDs and DVDs) and electronic storage in databases. Reprography is a form of reproduction, i.e. the duplication of a work. Reprographic reproduction is a process that usually results in a copy on a graphic surface by such process as printing, photocopying, scanning, digital copying (for instance on CDs and DVDs) and electronic storage in databases. The exact definition of reprography varies from country to country. As the process involves making a permanent reproduction of the work, it necessarily triggers one of authors’ exclusive rights – the reproduction right. For the researcher to be able to carry out reprography of copyrighted materials lawfully he or she has to obtain a permission from the relevant rightholder or the act of reprography needs to fall within the scope of a copyright exception. An exhaustive list of copyright exceptions is stipulated in the InfoSoc Directive.

Usually, the act of reprography is covered by so-called ‘private copying exception’ that is stipulated in Article 5(2)(b), which reads as

“in respect of reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the rightholders receive fair compensation which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned”

It should be noted that the provision utilizes the term ‘natural person’ to specify beneficiaries that can benefit from the exception. The InfoSoc Directive does not provide a definition for ‘natural person’. Arguably, a ‘natural person’ implies an individual human being, as opposed to a legal person, which may be a private business entity or public organisation.

For the researcher to be able to benefit from the private copying exception two conditions have to be met. First, the reproduction of the copyrighted work is used privately for ends that are neither directly nor indirectly commercial. Second, the rightholder receives fair compensation. The latter condition is straightforward and often requires little or no effort on the side of the researcher as

---

23 the International Federation of Reproduction Rights Organisations (2010)
many countries have a system of copyright levies, which ensures that rightsholders receive an appropriate compensation. The system relies on a tax that is charged on purchasing devices that enable copying. The money collected in as tax is then distributed among rightsholders. The first condition is complex and depends on many factors. Regarding nature of use, ‘private use’ implies that that the work, which is subject to reprography, has been lawfully acquired. In France, a student was exempted from liability despite been caught with 488 CDs containing variety of motion pictures that were copied from films he borrowed from friends.\textsuperscript{24} Hence, not all Member States have set the number of permissible copies as a prerequisite for the permissibility of private copying. In addition, the notion of ‘lawfully acquired’ should be understood broadly to include renting and borrowing.

Although, it might be challenging for the researcher to determine non-commercial nature of use because he or she has to decide whether any indirect economic gains result from such use, it is still conceivable that reprography for purposes of TDM may fall within the scope of private copying copyright exception. For example, if the researcher performs reprography on copyrighted paper sources that he bought or borrowed from a library, or a friend, i.e. obtained lawfully, such activity will be regulated by private copyright exception and considered lawful. Worth noting that, in practice, researchers almost never are engaged in reprography themselves, and consequently they rarely have to rely on private copying copyright exception in the given context.

Despite the fact that digitalizing of analog sources creates a crucial precondition for TDM activities, any legal issues that may arise from reprography are not part of the legal uncertainty that surrounds TDM. Digitalizing is a costly process because it requires manual work. Putting the burden of digitalizing on the researcher is likely to render the technology impractical for application because the researcher will either not able to conduct digitalizing at such a big scale or a drastic increase in cost of a research will force the researcher to resort to conventional research practices. In practice, sources for TDM are either initially created in the form of digital materials or they are digitalized by rightholder himself.

Scholars describe practical implementation of data analysis differently. Reviewing several approaches will allow me to identify the steps that, although not present in every instance of data analysis, are still representative for the technology as they usually take place during the course of a technology-enabled research.

\textsuperscript{24} Karapapa (2012) p. 94
Guadamuz and Cabell in their article ‘Data mining In UK higher education institutions: law and policy’ argue that data analysis consists of the six following steps:25

(i) Individual content is created
(ii) Content is placed into data set, repository or collection
(iii) Miner gains access to the data
(iv) Mining tools applied to the data set
(v) Analysis of the processed data
(vi) New knowledge

Steps (i) and (ii) do not constitute technological processes that comprise the phenomenon as they happen prior to inception of data analysis and do not require active actions on the side of the researcher.26 It is not clear what authors include in step (iii) ‘gaining access to the data’ because they give very little description, but I assume that it should include at least three processes:

(1) finding large quantities of the relevant materials,
(2) obtaining copyright authorization from copyright holders for part of the materials, which is copyrighted and
(3) performing an act of reproduction by making a copy of materials in non-volatile memory.27

It is defensible to say that the authors recognize step (iv), inter alia, on the ground that it is concerned transforming the previously obtained data into processed data. For example transforming DOCX to plain text. It should be note that researchers may skip step (iv) by limiting sources of data analysis to the data that belong initially to certain format. For instance, the researcher chooses to copy only those materials that are in DOCX format, but disregards materials formatted in PDF because the TDM software can work with DOCX immediately, while PDF needs to be converted to DOCX. Limiting sources negatively affects outcome of data analysis because the findings of the research will be based on fewer evidences. Step (v) Analysis of the processed data is an indispensable stage of a technology-enhanced research. Finally, step (vi) can hardly be considered as a valid stage because it denotes the outcome rather than a process of data analysis.

Scholars Maarten Truyens and Patrick Van Eecke in their article ‘Legal aspects of text mining’ suggest another approach that includes two phases such as ‘creation of corpus’ and ‘use of corpus’,

25 Guadamuz and Cabell (2014) pp. 5-6
26 Usually, an author creates the content and a publisher places the content into data set, repository or collection.
27 Non-volatile memory is computer memory that can retain the stored information even when not powered.
the latter is not relevant for the current discussion as it is focused on the applicable legal regime once a corpus is created. The term ‘corpus’ means a collection of texts\textsuperscript{28,29}. The creation phase includes steps such as

(i) Finding relevant texts
(ii) Preparing texts for inclusion in a corpus
(iii) Copying texts into corpus

As none of the steps appears to include analysis as its process, it is reasonable to conclude that implementing all the steps will result in creation of a corpus that consists of uniformly formatted texts rather than new knowledge. Introducing a stage similar to step (v) ‘analysis of the processed data’ of the algorithm above could improve this approach. Considering step (iii) ‘copying texts into corpus’, it should be noted that researchers can carry out data analysis directly from the publisher’s database using an API provided by the publisher on his website.

Jean-Paul Triaille in his ‘Study on the legal framework of text and data mining (TDM)’ (the study of Triaille et al.) suggests a generic TDM model that includes five steps such as

(i) Obtaining of the sources
(ii) Transformation of the data to fit operational needs
(iii) Loading of the data
(iv) Analysis of the data
(v) Drafting of a report

The study provides for a description for each step that is meant to help the reader to get better understanding of a process. As follows from the description, step (i) ‘obtaining of the sources’ to some extent corresponds to step (i) ‘finding relevant texts’ of the previous approach as it acknowledges that sources can be created from scratch or obtained from a third party.\textsuperscript{30} The authors’ use of the term ‘extraction’ varies. In one sentence,\textsuperscript{31} they use ‘extracting’ as ‘obtaining’, while in another sentence,\textsuperscript{32} ‘extraction’ means converting data into another format. The way the term ‘extraction’ is used is confusing as it denotes two distinct processes – obtaining and converting. It could benefit the current account to disaggregate the term ‘extraction’ into

\textsuperscript{28} The authors refer to text instead of data as they have their focus narrowed to text mining (data analysis based solely on text sources). Referring to text makes no difference for the present discussion as it endeavors to identify generic processes that are common for data analysis as a technology.
\textsuperscript{29} Truyens and Van Eecke (2014) p. 154
\textsuperscript{30} Triaille et al. (2014) p. 46
\textsuperscript{31} ‘It usually involves extracting the data from the source systems but a direct access to the sources is also conceivable’. Ibid. p. 45
\textsuperscript{32} ‘[T]he goal of an extraction […] is to convert the data into a single format appropriate for transformation Processing’. Ibid.
‘obtaining’ and ‘converting’ with moving the latter to step (ii) ‘transformation of the data to fit operational needs’ as it prima facie appears to have its goal as data transformation. Another disadvantage using ‘extraction’ to denote a technological process comes from the fact that term ‘extraction’ is likely to be understood within the meaning of Article 7(2)(a) of the Database Directive) which reads as

“‘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”.

Using another term such as ‘obtaining’ might help to avoid confusion. To merits of this account, authors state explicitly that data analysis does not always require obtaining digital materials via copying as accessing the sources directly33 is also conceivable.34

Application Programming Interface (API) allows researchers to carry out data analysis directly from the publisher’s database. API is a way for two computers to communicate with each other without either one having to know anything more than the XML.35 As accessing digital materials via API allows the researcher to avoid copying, this form of access can solve the legal uncertainty around TDM. Nevertheless, at least two factors prevent API to become a technical solution for legal issues that surround TDM. First, not all publishers have implemented API on their websites. Second, API does not allow the research to access digital materials of different publishers. As API has serious technical limitations, conducting of a substantial study with many sources is likely to require the researcher to rely on extensive copying of digital content.

When considering step (i) ‘obtaining of the sources’, it should be said that at this stage the relevant data is found, and then copied. For the researcher to be able to copy copyrighted materials lawfully he or she has to obtain a copyright permission. It is normal to request permission via the publisher of the work. The publisher will often have a permissions department to deal with such requests, or may use the services of a copyright clearance or licensing service. If the publisher cannot give permission directly, they will certainly know who the researcher should contact, (as they have obtained permission themselves in order to use the work in the first place). For website content, it is normal to contact the webmaster of the site. The webmaster may either give permission directly or refer the researcher’s request to someone in the company who can deal with it. A copyright permission is granted against paying a certain fee. If researcher’s intention is to make a

33 API is one way to implement direct access to digital materials. For purposes of this paper, direct access means a situation where a researcher does not need to copy the content to a separate location to conduct data analysis.  
34 Triaille et al. (2014) p. 46  
35 Extensible Markup Language (XML) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable.
commercially salable product, he or she should normally expect to pay royalties linked to the number of units sold. For non-commercial products, a flat fee may often be negotiated. When applying for a copyright permission it is advisable to give a full description of the work the researcher wish to use and specify how the work will be used.

The requirement for authorization in relation to copyrighted works is stipulated in Article 2 of the InfoSoc Directive and reads as

“Member States shall provide [authors] for the exclusive right to authorize or prohibit […] reproduction by any means and in any form […] of their works”.

Copyrighted works are often incorporated into databases. The Database Directive grants database makers or owners *sui generis* database rights. The reason for granting these rights is that databases makers should be able to recover what they invested in creation of databases. *Sui generis* database rights are stipulated in Article 7(1) of the Database Directive, which reads as

“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”.

In addition to *sui generis* database rights, database makers or owners may also be granted copyrights for their databases. However, only so-called ‘creative databases’ are granted copyrights. A database that involves a creative selection and arrangement of materials meets the requirement of creativity and may be eligible to receive *sui generis* and copyright protection. ‘Creativity’ depends on creative freedom and free and creative choices. The ECJ has defined boundaries of creativity in a negative way by stating that where choices are dictated by technical function, rules or constraints, the author is not able to “[express] his creative ability in an original manner by making free and creative choices”.36 In the databases context, it means that a database in which the content has been selected and arranged based on the optimization rationale or other technical considerations is less likely to receive copyright protection. Article 3(1) of the Database Directive, which grants copyright protection to ‘creative databases’, reads as

36 Football Dataco Ltd and Others v. Yahoo! UK Ltd and Others, par. 38
“[...] 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. No other criteria shall be applied to determine their eligibility for that protection”.

Copyrights and *sui generis* rights in relation to a database can overlap, but they are granted and exist independently. Works that are incorporated into a database have also their own copyright regime, which does not depend on whether the database is protected by copyright, this follows from Article 3(2) of the Database Directive, which reads as

“The copyright protection of databases provided for by this Directive shall not extend to their contents”

It is important to distinguish between extraction of a work and extraction of a database. Article 7(2)(a) defines ‘extraction’ as “the permanent or temporary transfer of all or a substantial part of the contents”. The term ‘substantial part’ is crucial for the notion ‘extraction’ because not every act of copying from a database is considered to be ‘extraction’ and triggers *sui generis* database rights of a database owner. If the amount of works that the researcher wants to copy from a database is small\(^\text{37}\), then the researcher has to obtain a copyright permission from the copyright holder for every work he or she wants to copy. \(^\text{38}\) However, if the number of works the researcher wants to copy from a database is considerable, i.e. constitutes a ‘substantial part’ of the database, then the researcher has to seek a permission for the database extraction and a copyright permission for each of the works he or she is intended to copy. Extraction of a database can be viewed as an extensive or bulk copying of database contents that triggers *sui generis* rights. In comparison, extraction of a work is an act of selective copying of database contents that triggers copyrights.

Moving back to the TDM technological model suggested by Triaille et al., it should be said that step (i) ‘obtaining of the sources’ or a similar stage in other models is closely related to the notion ‘level of access’. This notion is used to delimitate various types of content sharing depending on the number of participants and the legal basis for such sharing. The study of Triaille et al.\(^\text{39}\) recognizes four levels of access such as

(i) ‘All to all’ access level covers ‘web data’ that can be found on the websites, which impose no contractual terms upon users.

(ii) ‘Many to many’ access level covers ‘social network data’ that can be found on social networks, use of such data is limited by account settings and social network’ terms of use.

\(^{37}\) Small in relation to the total number of works that is stored in the database.

\(^{38}\) In practice, database owner and copyright holder of the works that are stored in a database is the same entity.

\(^{39}\) Triaille et al. (2014) pp. 18-19
(iii) ‘One to many’ access level covers ‘publishers’ data’ that is provided by publishers and/or repositories, use of such data is restricted by contractual clauses.

(iv) ‘One to one’ access level covers ‘confidential data’ that is disclosed by one entity to another entity under a confidentiality agreement.

Step (i) ‘obtaining of the sources’ should be qualified as an integral and essential part of the technology as it creates necessary conditions for further technological processes. The step can be implemented via copying or accessing digital contents directly via an API.

If implemented via copying, it triggers the reproduction right provided in Article 2 of the InfoSoc Directive. In case if works are copied from a database it may trigger a sui generis database right granted to a database owner by Article 7(1) of the Database Directive in addition to the reproduction right.

The next step (ii) ‘transformation of the data to fit operational needs’ involves series of functions that are applied to the data that has been obtained via copying at the previous stage.\textsuperscript{40} The rationale behind this step is to meet the technical requirements of data analysis software. Data manipulations vary depending on scale and complexity of transformation involved. A simple manipulation might involve converting PDF to XML or performing a simple mapping by replacing certain words or characters in the original text. From the copyright perspective, converting digital material from one data format to another is likely to be considered as reproduction. Other manipulations include sorting, joining data from multiple sources and de-duplicating the data, aggregation, etc. In essence, this step deconstructs the human language of text and reconstructs it for machines.\textsuperscript{41}

The next step (iii) ‘loading of data’ is another step in the TDM technological model suggested by Trialle et al. The only purpose of step (iii) is to load data into the end target. The study explains that end target can be a server, a hard disk or a data warehouse.\textsuperscript{42} In comparison, step (i) includes finding the relevant data, obtaining a copyright and/or sui generis database permission and downloading, i.e. loading data from the Internet. The scope of step (iii) is included into the scope of step (i), and therefore it is not clear why the authors stipulate this stage (iii) as it in essence repeats step (i) ‘obtaining of the sources’. Another argument against this stage is that, usually, step (iii) ‘loading of data’ precedes step (ii) ‘transformation of the data to fit operational needs’ as step (iii) creates necessary preconditions for step (ii). Redundancy and breaching of causal loop undermine validity of this stage of data analysis.

\textsuperscript{40} Step (ii) ‘transformation of the data to fit operational needs’ is not applicable for data analysis with direct access as data has been already transformed by the publisher to meet the needs of technology-enhanced research

\textsuperscript{41} Clark (2013) p. 12

\textsuperscript{42} Trialle et al. (2014) p. 48
Step (iv) ‘analysis of the data’ is a complex step that involves many processes. At this stage, the data that was obtained and transformed at the earlier stages is analyzed. Common processes at this stage include, but not are limited by, such tasks as: writing analytical memos, developing and applying a coding schema, retrieving of coded segments, hyperlinking, mapping, generating output, etc.\(^\text{43}\) Triaille et al. argue that this step can involve acts of copying. No doubt, the researcher can always decide to enrich further the corpus by carrying out some extra copying of digital materials. This reasoning is applicable to any stage of data analysis and, consequently, the reproduction right may be triggered at any stage. However, it is likely that all necessary copying has been performed at earlier stages when obtaining sources and creating the initial corpus. If any copying happens, it is likely to be incidental as this stage is mainly concerned with conducting analysis per se.

The last stage that is recognized by the study is step (v) ‘drafting of a report’. Common tasks of data analysis suggested by Silver and Lewins includes such a task as ‘generating output’. Considering the relevant descriptions of the stage\(^\text{44}\) and the task\(^\text{45}\), it is defensible to say that ‘drafting of a report’ is a process that result in generating output in the form of a final report. Task ‘generating output’ intrinsically belongs to step (iv) ‘analysis of the data’ as no analysis can be deemed complete without generating some sort of an outcome. From the formal point of view, excluding ‘generating output’ and elevating it to an independent stage would be logically inconsistent because it would imply the researcher taking on the next step of data analysis without fulfilling the previous one. This logic makes me argue against recognizing step (v) ‘drafting of a report’ suggested by the study.

2.2 Definitions of the term TDM in the literature

Many researchers have tried to define the concept of TDM. Some of these definitions provide a good description of various aspects of the technology. Few definitions that are found in the literature give the holistic account of the activity by focusing on its most essential features. In this section, I will review some of the recent definitions suggested by legal commentators, research institutions and publishers. Considering different accounts of the phenomenon should allow me to identify items that pertain to a sound TDM definition. Findings made in this section will be used in the next section when assessing enacted and draft legislation on TDM. Getting ahead of myself,

\(^{43}\) Silver and Lewins (2014) pp. 9-10
\(^{44}\) ‘The production of an output […] which is readable by humans’, see Triaille et al. (2014) p. 49
\(^{45}\) ‘Report on […] the project […] to get away from the computer and think and work in more “traditional” ways’, see Silver and Lewins (2014) p. 10
I need to say that the expression TDM, for the reasons explained below, is imprecise. However, because the acronym is often used in literature and legislation, it may surface later in the text.

Definitions reviewed below do not refer to the fact that TDM activities might lead to reproduction or creating derivative works, which I consider as an advantage because a stable definition should avoid having build-in legal terminology. The definitions are not assessed in relation to copyright, but from the point of view, how well it captures factual dimension of the phenomenon.

Jonathan Clark argues that

“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”\(^{46}\)

This definition holds that the difference should be made between ‘text mining’ and ‘data mining’. Considering the correlation between ‘text’ and ‘data’, it should be said that the EU lawmakers have given a broad definition of ‘data’ in the Database Directive.\(^{47}\) A database is by definition composed of ‘data’, but, in reality and in accordance with the Database Directive, it contains materials such as text, sound, images, etc.\(^{48}\) ‘Data’ should therefore be considered as a generic term that includes all types of content such as text, images, videos, and many others. Consequently, ‘text mining’ is just one subset of ‘data mining’. If choosing among ‘text mining’, ‘text and data mining’ and ‘data mining’, the latter term should be preferred. The wording ‘turns text into data’ is controversial since text in a formal sense is a form of data, albeit unstructured. The phrase ‘analytical process’ in singular belittles the phenomenon. According to Silver and Lewins, analysis is a series of processes that are fluid and overlapping.\(^{49}\) Data mining is computational analysis,\(^{50}\) and therefore should be viewed as a series of processes.

The study of Jean-Paul Triaille provides for an informed account of the phenomenon by suggesting the following definition:

---

\(^{46}\) Ibid. pp. 5-6


\(^{48}\) Recital 17 of the Database Directive, reads as ‘[…] the term “database” should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sounds, images, numbers, facts, and data’.

\(^{49}\) Silver and Lewins (2014) p. 16

\(^{50}\) Tay and Sik (2016) p. 6
“data analysis [i]s [t]he 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”.51

Comparing the terms ‘data analysis’ and ‘data mining’, it should be noted that ‘to mine’ means extracting knowledge from data sources similarly to extracting minerals from the earth.52 The term ‘data mining’ emphasizes ‘extraction’ as a research activity, but the phenomenon involves many interrelated and fluid processes happening simultaneously. Using the term ‘mining’ create a misleading impression that the phenomenon is limited to ‘extraction’. In addition, as correctly observed in the study,53 it is often understood that ‘to mine’ content means ‘to go deep into’ text, videos, images, etc., whereas some instances of the analysis might involve only techniques, such as crawling and scraping, that ‘stay on the surface’ of content. Crawling is an automated searching through a website and scraping is an automated downloading of substantial parts of a website.54 The expression ‘automated’ expands the scope of this definition beyond computer technology, and thereby makes it more technology neutral. As reasonably suggested in the study, there may, and in the future probably will, be other automated techniques than using computer technology.55

‘Processing’ is a broad term that accommodates well both the methodological and the technological dimensions of the phenomenon. From a methodological point of view, ‘processing’, unlike a ‘process’ in singular, might imply that the activity involves several analytical processes occurring at the same time. From a technological point of view, the wording ‘processing’ acts as a non-exhaustive list of techniques that might include both text-specific and/or general methods that are relevant for all data types. By listing them together, separated by commas, the wording – ‘may include texts, sounds, data, images or other elements’ – implies that these notions are somewhat equal. However, the term ‘data’ is a generic category that also encompasses other terms.

Sergey Filippov and Paul Hofheinz claim that

“text and data mining [i]s an advanced algorithm-based reading method in which large volumes of text and data from existing articles can be analysed, categorised and sorted to detect patterns and extract meaningful information for a wide variety of purposes”.56

51 Trialle et al. (2014) p. 17
52 Borghi and Karapapa (2013) p. 47
53 Ibid. p. 10
54 Jennings and Yates (2009) p. 120
55 Ibid. p.17
56 Filippov and Hofheinz (2016) p. 2
Referring to the phenomenon as ‘text and data mining’ is not fully correct because the term ‘data’ includes the term ‘text’. The term ‘method’\textsuperscript{57} is too narrow as the activity may involve several techniques, such as classification, clustering, prediction and many others, at once.\textsuperscript{58, 59} The wording ‘text and data’ is questionable because, as it has been discussed above, the expression ‘data’ is a more general term that includes ‘text’. The expression ‘articles’ is limiting as data also comes from other sources than articles. The wordings ‘analysed, categorised and sorted’ and ‘detect […] and extract’ do not provide for a complete list of the techniques available.\textsuperscript{60} The wording ‘for a wide variety of purposes’ benefits the definition by providing it with a non-exhaustive list of purposes.\textsuperscript{61}

According to the UK National Centre for Text Mining,

“[t]ext mining is the process of discovering and extracting knowledge from unstructured data. This comprises three main activities: (i) [i]nformation retrieval (IR) to gather relevant texts; (ii) [i]nformation extraction (IE) to identify and extract entities, facts and relationships between them; (iii) [d]ata mining to find associations among the pieces of information extracted from many different texts”.\textsuperscript{62}

This definition is limited as it refers to ‘the process’ and it relies on an exhaustive list of techniques. Moreover, data mining is subjugated to text mining here, whereas, as also has been explained previously, the opposite is true. ‘Unstructured data’ is limiting because structured data may also be processed and analyzed. The purpose of the analysis – ‘discovering and extracting knowledge’ - is formulated inadequately. While ‘discovering’ has a sufficient level of generality that makes the definition applicable to instances of analysis with unforeseen purposes, ‘extracting’ would be more appropriate in the context of methods or techniques rather than objectives.

Another definition comes from a sample standard text and data mining license drafted by the International Association of Scientific, Technical & Medical Publishers (STM). It is stated in the license sample that “‘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

\textsuperscript{57} According to Merriam-Webster dictionary, the terms ‘method’ and ‘technique’ are synonyms because it defines ‘method’ as ‘a systematic procedure, technique, or mode of inquiry’.

\textsuperscript{58} Brown (2012)

\textsuperscript{59} A broader term that encompasses a variety of methods would be more apt here. For instance, the authors could use ‘technology’ instead of ‘method’.

\textsuperscript{60} For example, the definition that has been suggested by the International Association of Scientific, Technical and Medical Publishers and will be considered later in this section acknowledges more forms of analytical inquiry such as ‘the sorting, parsing, addition or removal of linguistic structures, […] the selection and inclusion […]’.

\textsuperscript{61} Indeed, it is hardly possible to envisage all the purposes for which the technology may be used.

\textsuperscript{62} The UK National Centre for Text Mining
recognition of relations and associations”\textsuperscript{63} The expression ‘automated’ makes this definition more technology-neutral as it does not limit the activity to computer technologies.\textsuperscript{64} The wording ‘Publisher’s Content’ is beneficial because it is a broad term that covers all sorts of works or data. Although the list of techniques provided in this definition is incomplete as an experienced computer scientist can name a method that is not on the list, so far it is the most comprehensive account of available techniques. The purposes of the activity – ‘classification or recognition of relations and associations’ – are formulated as a further listing of techniques, leaving this account of phenomenon with no sound definition of objectives. Referring to the activity as TDM makes this definition less accurate, comment above.

3 On the current exceptions

Scholars such as Ian Hargreaves, Jonathan Clark and Sergey Filippov argue that researchers face legal uncertainty when conducting data analysis. As explained previously in section 1.3, the legal controversy that surrounded machine reading that precedes data analysis came from the fact that copies made for machines could not be comprehended by humans, and therefore appeared to be outside of the copyright scope. It is evident from the discussion above on data analysis technological processes that the technique of obtaining data results in copies that can be understood by humans. Moreover and what is more important, such copies retain the expressive value of the original work they were taken from. These facts indicate that the legal ambiguity that inhibits data analysis differs from the one that was discussed in section 1.3. In this chapter, I will examine data analysis through the prism of the existing exceptions to IPRs. Doing so should allow me to decide whether the technology falls within the scope of any existing limitation to IPRs.

3.1 Copyright law

The InfoSoc Directive includes copyright exceptions that possibly may cover TDM activates. One of such exception has been considered earlier in the beginning of section 2.1 – private copying exception. However, the exception is inapplicable for purposes of data analysis. Another candidate

\textsuperscript{63} The International Association of Scientific, Technical and Medical Publishers (2012) p. 1

\textsuperscript{64} In contrast, the account of the phenomenon that is given in Japanese legislation and will be analyzed in the next section uses wording ‘by computer’.
– scientific research exception – is stipulated in Article 5(3)(a) of the InfoSoc Directive and is meant to benefit scientific research.\(^{65}\) The article reads as

“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”.

‘Illustration’ is an important notion for research exception as it constitutes the purpose of the activity that can benefit from the exception. ‘Illustration’ is not defined anywhere in the InfoSoc Directive. In its ordinary meaning, illustrate means “to clarify something by giving or serving as, an example of a comparison”\(^{66}\). In a research context, ‘illustration’ could be understood as allowing the researcher to reproduce or otherwise use a work “as an example”. Under data analysis, works are not used as an example, therefore ‘illustration’. Hence, ‘illustration’ creates an obstacle for applying the research exception to data analysis.

Another obstacle in applying the scientific research exception to data analysis stems from the requirement to the nature of the activity. Recital 42 of the InfoSoc Directive provides that scientific research is a non-commercial activity. The recital reads as

“When applying the exception or limitation for noncommercial educational and scientific research purposes, including distance learning, the non-commercial nature of the activity in question should be determined by that activity as such […]”

The recital mentions the organisational structure and the means of funding of the establishment among the factors that should be taken into account. Nevertheless, the decisive factor for determining the non-commercial nature of the activity is ‘the activity as such’. The wording ‘the activity as such’ is confusing and difficult to penetrate. The study of Triaille et al., suggests five cases where a research funded by a private company may still be considered having a non-commercial nature. However, it is acknowledged in the study that the commercial v. non-commercial criterion is hard to apply.\(^{67}\) The wording of the article is likely to exclude the majority of mixed projects that include academic and commercial entities from the scope of the research exception. The exemption mandates that the source of each copyrighted work must be indicated unless this turns out impossible. Considering that success of a research based on data analysis to a certain degree depends on amount of works it used, it could be difficult for researchers to give an

---

\(^{65}\) Article 5(3)(a) of the InfoSoc Directive reads as ‘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’

\(^{66}\) The online Merriam-Webster dictionary

\(^{67}\) Ibid. 67
exhaustive list of attributions. The research exception states that the use of a work is limited ‘to the extent justified by the non-commercial purpose’. In other words, the amount taken from the work should not exceed what is required by the researcher. It could be difficult for the researcher to fulfill this requirement, as normally s/he does not know what amount of work is relevant for the research and whether the work has any value for the research at all. Considering millions of works used for a single research, it is likely that at least some of them do not contribute to research as they were copied by accident (an article includes the relevant keywords, but talk about absolutely different subject). In another case, while the research downloaded the whole article, but only a small part it was consequently used. It means that amount taken will always exceeds what is necessary. Finally, the implementation of the scientific research exceptions varies among Member States to with respect to the types of institutions covered and the length of the excerpts that can be taken from the work. Hence, data analysis is unlikely to be covered by the exception.

The temporary reproduction exception is another copyright limitation that is relevant for data analysis.  

The exception is stipulated in Article 5(1) of the InfoSoc Directive, which is read as

“Temporary acts of reproduction referred to in Article 2, which are transient or incidental [and] an integral and essential part of a technological process and whose sole purpose is to enable:

(a) a transmission in a network between third parties by an intermediary, or

(b) 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 exemption is formulated as a set of cumulative conditions, meaning that non-compliance with any of them will prevent an act of copying benefit from the exception. It follows from the wording of the provision that only transient or incidental acts of reproduction are covered by the exemption. Terms ‘transient’ and ‘incidental’ were considered by the ECJ in cases C 5/08 (Infopaq I) and C-302/10 (Infopaq II). According to the court, the duration of a transient copy is limited to what is necessary for the proper completion of the technological process that enables browsing of the copyrighted work. While the decision does not set an absolute lifespan limit for a transient copy, its wording – ‘not exceed what is necessary’ – suggests that the duration of a transient copy must be limited. Another criterion established by the court is absence of human intervention in

68 The InfoSoc Directive, Art.5(1)
69 Infopaq International A/S v. Danske Dagblades Forening (2009), par. 33, 64
70 Ibid., par. 61
processes such as creation and deletion of a transient copy. However, the court reconsidered its stance on human intervention in Infopaq II by stating that a technological process that involves human interaction could result in a transient copy. Copies created when obtaining sources for data analysis are unlikely to be considered transient as they neither serve the purpose of enabling browsing of the work nor they cease to exist once data analysis is finished.

The article also recognizes incidental acts of copying. According to Merriam-Webster's Collegiate Dictionary, ‘incidental’ means occurring merely by chance or without intention or calculation. Taking into account the ordinary meaning of ‘incidental’ it is reasonable to assume that incidental copying is primarily concerned the intention of the person who performs copying. Nevertheless, legal content of a term can deviate strongly from the ordinary meaning of words it involves, and therefore it would be unwise to rely solely on it. Fortunately, ‘incidental’ caught attention of the ECJ in case Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others. The court considered relation between ‘transient’ and ‘incidental’ by comparing on-screen and cached copies. In contrast to on-screen copies, cached copies do not cease to exist when the user terminates the technological process used for viewing the website. Despite the fact that incidental copies last longer than transient ones they still cannot be permanent as it would contradict ‘temporary’ that is used earlier in the sentence. As none of the processes discussed in section 2.1 involve manually activated or performed automatically deletion process of works, copies made during the course of data analysis cannot be considered incidental. In the light of the discussion above, it is logical to conclude that copies made for data analysis are neither transient nor incidental and therefore the condition regarding the temporary nature of reproduction would not be met. Failing to satisfy one of the conditions stipulated in article voids the application of the exception because the conditions are cumulative.

Another limitation that has to be examined is quotation exception. The exemption permits copying from a copyrighted work for purposes of criticism or review. Similar to scientific research exception, quotation exemption requires including author’s name. The attribution condition is to some extend is mitigated by the fact that it can be skipped in cases where it turns out to be very difficult or even impossible to indicate all authors’ names due to big number of works have been used. Despite the relaxed attribution condition, the quotation exception is of limited use for the process of obtaining works as both the nature of a quotation and the wording of the exception

---

71 Ibid., par. 63
72 Infopaq International A/S v. Danske Dagblades Forening (2012), par. 32, 36, 39
73 Public Relations Consultants Association Ltd v. Newspaper Licensing Agency Ltd and Others (2013)
74 The first sentence of the InfoSoc Directive, Art.5(1) reads as ‘[t]emporary acts of reproduction referred to in Article 2, which are transient or incidental […]’.
75 The InfoSoc Directive, Art.5(3)(d)
assume that only a few fragments of a work can be lawfully used. As data analysis thrives of quantity approach the selective approach towards copying stipulated in the exception intrinsically contradicts to the underlying principle of the technology.

After considering three exceptions that were most likely candidates to cover acts of reproduction of works that are involved in data analysis processes it became evident that they either explicitly inapplicable or their application entails high level of legal uncertainty.

Case law may benefit the assessment of copyright exceptions’ applicability to data analysis. As correctly observed by legal commentators Truyens and Eecke case law on data analysis is sparse. Nevertheless, some cases still can be found, among them the case that was brought before the District Court of Amsterdam by the Swiss Anne Frank Fonds against the Dutch Anne Frank Stichting and the Royal Netherlands Academy of Arts and Sciences. In that case, XML files were created from Anne Frank’s manuscripts and published diaries to make a textual analysis of those works. The Swiss Anne Frank Fonds, which owns copyrights in these manuscripts and diaries, did not authorize the creation of the XML files, and therefore claimed that they infringed its copyrights. The defendants invoked several copyright exceptions, inter alia, quotation exception that is stipulated in the Dutch law. The court rejected the line of defense based on asserting that the reproductions were made under the quotation exception stipulated in Article 15a of the Dutch Copyright Act. According to the court, while taking a portion of a work is generally permissible under the exception, the XML files in question contained the full text of the work. Admitting that reproduction of an entire work is permitted under specific circumstances, the court nevertheless held that the defendants failed to prove any facts that would justify bulk copying. The defendants tried to pursue other lines of defense evoking various exceptions, but the court found none of the exceptions applicable to justify the bulk reproduction of works. Fortunately, for the defendants they also invoked their freedom of scientific research – as laid down by Article 13 of the EU Charter of Fundamental Rights (the EU Charter). The court assessed the between the enforcement of IPRs and fundamental rights granted by the EU Charter and decided that the freedom of scientific research outweighed the enforcement of copyrights. Justifying its conclusion the court stated that as only a few researchers have access to only a few copies of Anne Frank’s diary the infringement had no more than minimal impact.

---

76 Truyens and Eecke (2014) p. 169
77 Anne Frank Fonds v. Anne Frank Stichting and the Royal Netherlands Academy of Arts and Sciences, par. 4.7.1.
78 Ibid.
79 Ibid. par. 4.8.1.
80 Ibid. par. 4.8.3.
81 Ibid.
reinforces the finding made earlier in this section – data analysis is not covered by any of the exceptions stipulated in the Infosoc Directive.

3.2 Database law
Since data analysis obtains data from databases, the EU database legislation becomes a relevant set of rules for the technology. If you look at data analysis via a prism of database law, TDM involves an act of extraction. Article 7(2)(a) of the Database Directive defines extraction as “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”. Article 9 of the Directive lists three exceptions, but only one of them – the scientific research exemption stipulated in Article 9(b) – is relevant for data analysis.

The article reads as

“In the case of extraction for the purposes 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”

Similar to the InfoSoc Directive, the Database Directive relies on the notion ‘illustration’ when formulating its research exception. The Database Directive provides no definition for ‘illustration’, which makes me to rely on the ordinary meaning of the term – “an example or instance used to make something clear”. Can a copy of a substantial part or the whole database been considered as an example? I do not think so because in my understanding, ‘an example’ is something short and precise meant to clarify a certain issue, while a copy of a database is so complex by itself that it can hardly clarify anything before it gets interpreted itself in the first place. Hence, the term ‘illustration’ does not support an application of the exception to data analysis.

Worth noting that the scientific research exception covers extraction, but it does not cover re-utilization. The act of re-utilization is defined as “any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies [or any] forms of transmission”. Fortunately, TDM use of databases does not contribute to their re-utilization because the output of TDM that is meant to made available to the public under no circumstances can be considered a copy of the utilized database. Similar to the scientific research exemption in

82 Article 7(2)(a) of the Database Directive defines ‘extraction’ as ‘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’
83 The online Merriam-Webster dictionary
84 The Database Directive, Art. 7(2)(b)
the Infosoc Directive, the exception stipulated in Article 9(b) has cumulative character. It includes four conditions:

(i) Lawful user  
(ii) Data are extracted for the purpose of scientific research  
(iii) The source is indicated  
(iv) The non-commercial purpose to be achieved

The term ‘lawful user’ is not defined in the Database Directive and it is controversial. The legal commentator Estelle Derclaye suggests that the lawful user can either be (i) ‘the user relying on statutory or contractual exceptions provided by law or by contract’, (ii) ‘the licensee’, or (iii) ‘the lawful acquirer’.\(^{85}\) In the light of Derclaye’s interpretation, ‘lawful user’ is a rather certain and achievable condition in the context of data analysis. The researcher may fulfill this requirement by simply entering into a license agreement with the database owner.

As explained in the context of copyright, the condition ‘scientific research’ is closely linked to ‘non-commercial purpose’ since ‘scientific’ necessarily means ‘non-commercial’. Considering the overlapping character of copyright and database protection, it is reasonable to assume that the sui generis condition ‘scientific research’ has a similar interpretation as in the copyright law, meaning that only ‘purely’ non-commercial researches may benefit from the exception. Any form of involvement of a commercial entity casts a shadow over the non-commercial character of a project and is likely to deprive the research to benefit from the exception.\(^{86}\)

Similar to Article 5(3)(a) of the Infosoc Directive, Article 9(b) imposes the attribution condition requiring to indicate the source of the database. In contrast to the analogous copyright exception, Article 9(b) does not drop this condition by using wording – ‘this turns out to be impossible’. The requirement to mention the source of all the databases is likely to become an insurmountable technical obstacle for researches that involve the processing of hundreds or thousands of sources, which are databases. As the absence of reference to ‘unless this turns out to be impossible’ leaves however no room for interpretation researchers might have to abandon data analysis.

After reviewing current copyright and database exceptions that are relevant for data analysis, it is defensible to say that the current legal framework is not fully adapted to data analysis, and therefore a legislative intervention in form of a new exception to IPRs is necessary.

\(^{85}\) Derclaye (2008) p. 120

\(^{86}\) Few cases were suggested in legal commentary when the involvement of a commercial entity will not change non-commercial nature of the research. However, none of this cases has been tested in a litigation. See Triaille et al. (2014) p. 65
Assessing the adequacy

4.1 Japan

In 2009, Japan amended Section 5 of the Japan Copyright Act on ‘Limitations to Copyright’ and introduced Article 47(7) on ‘Reproduction, etc. for information analysis’. The article reads as

“[f]or 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 such 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. […]” 87

As explained by the Japanese Group in the AIPPI questionnaire, the rationale for this exception is that ‘copyright shall not be exercised against the act of reproduction or adapting a work on a recording medium for the purpose of information analysis to be conducted on a computer under certain conditions’. 88

Article 47(7) ends with the following sentence:

“[h]owever, an exception is made of database works which are made for the use by a person who makes an information analysis”.

It may be challenging to conceive the meaning of this part of the Japanese TDM exception because the legislator wrapped an exception inside another exception. Explaining the meaning of this wording, Trialle et al. compared it to a legal provision that allows the reproduction of books for educational purposes, but discards the exception for scholar books. If permitting copying of schoolbooks for educational purposes, it would conflict with a normal exploitation of the work, and unreasonably prejudice the legitimate interests of the author. 89 Using the same logic, the Japanese TDM exception discards its application to databases that are created specifically for purposes of data analysis. Such an approach, prima facie, benefits the exception as it ensures the legitimate interests of database owners. However, some publishers regard data analysis as the main future vehicle for interaction with their publications, and therefore they consider changing their

87 The Copyright Act 1970 (Japan), Chapter ii, Sec.5, Subsec.5, Art.47(7) unofficial translation from Japanese to English by Yukifusa Oyama et al. taken from website of Copyright Research and Information Center from Tokyo
88 AIPPI Japanese Group (2011) p. 9
89 Ibid.
publishing practices. For instance, STM, a big publisher, has indicated that it will be publishing in an adapted manner to facilitate data analysis.\textsuperscript{90} If endorsed by other major publishers, this development may undermine the application of the exemption.

The approach taken by Japanese lawmakers has a few strong points. It defines broadly the source of data analysis as it refers to ‘information’ that includes “languages, sounds, images or other elements”. Such wordings as ‘it shall be permissible to make recording on a memory […] of a work’ and ‘it shall be permissible […] to make adaptation […] of a work” increase legal certainty as they explicitly authorize reproduction and/or adaptation of the work.

The term ‘analysis’ implicitly includes a variety of analytical technics and could have been further reinforced by the phrase ‘other statistical analysis’ as the word other, prima facie, introduces a non-exhaustive list of techniques. However, the selected term ‘statistical analysis’ is merely a subset of quantitative research and, by using it rather than some more general term, the provision makes me wonder whether a qualitative research, the one what does not rely on any numbers, is covered by the exception.\textsuperscript{91} It is generally understood that quantitative research deals with numbers, statistics or hard data whereas qualitative data are found mostly in the form of words.\textsuperscript{92} In contrast to the text-centric definition that is given by the UK National Centre for Text Mining, the Japanese TDM definition is biased towards hard data.

The wording – ‘by using a computer’ – adopted in the article goes in strong contrast to the term ‘automated’ used by the International Association of Scientific, Technical & Medical Publishers in its license sample. As correctly pointed out by Maurizio Borghi and Stavroula Karapapa, ‘the Japanese […] exception has the merit of clarifying a number of potential legal uncertainties in the online environment, it has the disadvantage of being bound by state-of-the-art technologies. As soon as technology changes, the list might need to be updated’.\textsuperscript{93}

The Japanese legislator excludes raw data – items of information that are not protected by copyright - from the scope of the exemption, as the provision only refers to the recording or the adaptation of a ‘work’. The term ‘work’ stipulated in Article 10(1) of the Japan Copyright Act\textsuperscript{94} is defined in a similar vein as the term ‘protected work’ in Article 2(1) of the Berne Convention.

\footnotesize
\textsuperscript{90}Thorburn et al. (2017) p. 132
\textsuperscript{91}My suggestion is that the lawmakers could have simply omitted the term ‘statistical’.
\textsuperscript{92}Chui (2007) p. 48
\textsuperscript{93}Borghi and Karapapa (2013) p. 62
\textsuperscript{94}The Copyright Act 1970 (Japan), Chapter ii, Sec.1, Art.10(1) reads as ‘(1) As used in this Law, “works” shall include, in particular, the following: (i) novels, dramas, articles, lectures and other literary works; (ii) musical works; (iii) choreographic works and pantomimes; (iv) paintings, engravings, sculptures and other artistic works; (v) architectural works; (vi) maps as well as figurative works of a scientific nature such as plans, charts, and models; (vii) cinematographic works; (viii) photographic works; (ix) program works’.

28
Similar to the Berne Convention Article 2(8), Article 10(2) of the Japan Copyright Act\textsuperscript{95} grants no copyright protection to daily news and miscellaneous facts having the character of mere items of information. While such approach might be sufficient in Japan since there is no database protection, adapting a TDM exception in a similar wording in the EU would result in an unbalanced exception. The reason for it is that raw data incorporated into databases would be excluded from the scope of the exception, which in turn would seriously undermine the exception.

The Japanese definition refers to the activity as ‘information analysis’ rather than ‘data analysis’. In ordinary language, ‘data’ is often used synonymously with ‘information’. However, in informatics, where information is viewed as being dependent on data as its vehicle, the terms are deemed as having different scopes. The International Organisation for Standardization (ISO) defines ‘data’ as ‘a reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing’.\textsuperscript{96} In comparison, ISO refers to ‘information’ as “the meaning assigned to data by means of conventions applied to that data”. The latter ISO definition reflects an approach that is suggested by information theorist, Luciano Floridi and defines ‘information’ as ‘data + meaning’.\textsuperscript{97} Information is derivative to data in the sense that in order to generate a particular type of information, data must first be structured in a particular manner. The relation between data and information is also explained via the Data-Information-Knowledge-Wisdom hierarchy, which embodies an assumed linear rise in epistemic values from Data (low value) through to Wisdom (high value), each stage being reached on the basis of ever-higher degrees of reflection, organisation, and accumulated learning.\textsuperscript{98} It appears that ‘information analysis’ is a less accurate term, as it gives a false impression that the phenomenon becomes relevant at the epistemic level of ‘information’ and onwards, while it figurates prominently in various definitions of the analysis that discovers meaning of data.

4.2 The UK

In 2010, the UK Prime Minister commissioned a report to consider whether the UK intellectual property framework is sufficiently well designed to promote innovation and growth in the UK economy. The report was carried out and published by Professor Ian Hargreaves in May 2011. Among other findings, the Hargreaves report indicated that the UK Government should introduce

\begin{itemize}
\item \textsuperscript{95} The Copyright Act 1970 (Japan), Chapter ii, Sec.1, Art.10(2) reads as ‘(2) News of the day and miscellaneous facts having the character of mere items of information shall not fall within a term “works” mentioned in item (i) of the preceding paragraph”.
\item \textsuperscript{96} ISO (2015)
\item \textsuperscript{97} Floridi (2011) p. 20
\item \textsuperscript{98} Rowley (2007) pp. 163-164
\end{itemize}
an exception, namely to support text and data analytics that would allow technology-enabled uses of works in ways which do not directly trade on the underlying creative and expressive purpose of the work. After the report, the UK Intellectual Property Office (IPO) attempted to define the scope and the limitations of the future TDM exception. Thus, in December 2012, the Government response to the report provided the following definition: “[a]utomated analytical techniques such as text and data mining work by copying electronic information, for instance articles in scientific journals and other works, and analysing the data they contain for patterns, trends and other useful information”.

The wording of the definition indicates that ‘text and data mining’ is viewed as an instance of ‘automated analytical techniques’. Comparing the UK approach to the earlier discussed Japanese approach, it draws attention that the UK wording uses the term ‘automated analytical techniques’, while the Japanese version in the same context employs the term ‘computer’. As it was elaborated above, in relation to the STM definition of TDM, the term ‘automated’ has an advantage over the expression ‘computer’ as being more technology-neutral, and thus includes a broader range of techniques. Regarding sources of analysis, the definition mentions as an example only literary works such as ‘articles in scientific journals and other works’. In contrast to this rather restrictive example, the wording ‘electronic material’ is likely to accommodate all the sources that are relevant for data analysis.

Mentioning copying as an imperative part of data analysis might be a drawback of this definition, as this is not always the case. For instance, if data analysis is carried out with help of Application Programming Interface (API) it can be done directly from the publisher’s database, and copying the work is no longer required. API is a way for computers to communicate with each other. In the context of text mining and scholarly publishing, API usually refers to a Web API expressed in XML. Web APIs are a way of sharing content and data between two computers without either one having to know anything more than the XML language used in the API and its rules. A simple API might be one that returns the PDF of an article. I will come back to the issue of copying later when considering copyright and sui generis database rights affected by data analysis.

In 2014, by amending its Copyright, Designs and Patents Act of 1988, the UK became the second jurisdiction worldwide that explicitly recognized and permitted data analysis. Section 29A provides that

“(1) [t]he 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—

99 Hargreaves (2011) p. 8
100 The UK Intellectual Property Office (2012)
101 Clark (2013) p. 19
(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, and
(b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).

(2) Where a copy of a work has been made under this section, copyright in the work is infringed if—

(a) the copy is transferred to any other person, except where the transfer is authorized by the copyright owner, or
(b) the copy is used for any purpose other than that mentioned in subsection (1)(a), except where the use is authorised by the copyright owner”.

The last part of the provision that is relevant for the current analysis - Section 29A(5) - reads as

“[t]o 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”.

According to the exception, only the person who has ‘lawful access’ to the work is permitted to make a copy of the work. The UK IPO explains that ‘lawful access’ covers cases where researchers have the legal right to access a copyright work to read it. ‘Lawful access’ can be obtained via paying for a subscription to a journal or database or using material published under open licenses. Clarifying what has changed after the exception was enacted, the office states that the exemption allows researchers to perform data analysis on copyrighted material that they have lawful access to.102 The explanation that is given by the UK IPO covers only a certain type of ‘lawful access’. However, the content of ‘lawful access’ varies greatly depending on the level of access to material. As discussed earlier different levels of access presume different conditions of lawfulness and cover different type of data. The term ‘lawful access’ benefits the UK exemption for data analysis as it ensures that only those who comply with the conditions of the relevant level of access can benefit from the exception.

Unlike the earlier TDM definition that is given in the Government response to the Hargreaves report, the enacted provision refers to the phenomenon as ‘computational analysis’. By changing the wording of the provision from ‘automated analytical techniques’ to ‘computational analysis’, the UK lawmakers took a similar stance to the Japanese legislator in relation to the technology involved. ‘By computer’ was discussed in length earlier in this paper; it appears that the term

102 The UK Intellectual Property Office (2014) p. 6
‘computational’ has the same disadvantage associated with the term ‘by computer’, namely being bound by state-of-art technologies.

The wording ‘anything recorded in the work’ is apt to its task because it is broad enough to encompass various sources that serve as an input for ‘data analysis’. The scope of the exemption is restricted by the fact that ‘data analysis’ can be carried out only for ‘a non-commercial purpose’. In contrast, the Japanese lawmakers adapted a more liberal stance that permits data mining irrespective of whether it is carried out for commercial or non-commercial purpose. Criterion ‘commercial v. non-commercial’ is often used in copyright law. For example, the InfoSoc Directive delimitates the notion ‘non-commercial’ in article 5(2)(c) which reads as

“in respect of specific acts of reproduction made by publicly accessible libraries, educational establishments or museums, or by archives, which are not for direct or indirect economic or commercial advantage”.

Jean-Paul Triaille argues in his study that the wording ‘for direct or indirect economic or commercial advantage’ should be understood as an act not giving rise to a payment the amount of which does not go beyond what is necessary to cover the operating costs of the establishment.\textsuperscript{103} Despite being tailored to establishment, the article gives some insight into the concept of non-commercial.

Another provision that is more relevant for the analysis as it tackles non-commercial nature of \textit{an activity} rather than of \textit{an establishment} is found in recital 42 of the InfoSoc Directive. The recital stipulates that

“[w]hen applying the exception or limitation for noncommercial educational and scientific research purposes, including distance learning, the non-commercial nature of the activity in question should be determined by that activity as such. The organisational structure and the means of funding of the establishment concerned are not the decisive factors in this respect”.

The wording “the activity as such” provides little help understanding the concept of non-commercial as it \textit{per se} does not explain what factors identify non-commercial nature of the activity. Triaille in his study suggests the five following cases when a research funded by a private company may still be considered non-commercial.\textsuperscript{104}

\begin{footnotesize}
\footnotetext[103]{Jean-Paul Triaille et al. (2014) p. 64}
\footnotetext[104]{Ibid. p. 65}
\end{footnotesize}
(i) If the funding agreement between the founder and the founded research institution states that, the research institution will own intellectual rights on the results.

(ii) If the funding agreement between the founder and the founded research institution states that the research results are not confidential and may be made public.

(iii) If a private company finances research for philanthropic purposes.

(iv) Even if founded by a private company, a ‘fundamental research’\(^\text{105}\) would normally qualify as research with no commercial purpose because it does not have an objective to bring a new product on the market.

(v) In certain cases, research is being made where the researchers intentionally publish the results of their work so that it becomes part of the ‘prior art’ in the patent law sense, and therefore prevent anyone from patenting research at question on that basis.

However, even being provided with meaningful factors, it still may be tricky to identify non-commercial nature of the activity. For instance, commercial aspects that are remotely present in a non-commercial research may render the research as having a commercial nature. It also brings controversy when the results of a non-commercial research are valorized in a monetary way afterwards. I will come back to the issue of identifying non-commercial nature of the activity when analyzing the proposed EU TDM. For now, it is sufficient to conclude that commercial v. non-commercial criterion is hard to apply, meaning that there will be borderline cases.

Apart from non-commercial purpose, other conditions restrict the scope of permissible reproduction. Another requirement that is stipulated in the provision is that ‘the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)’. Arguably, the UK lawmakers included this requirement in order to align their new exception with the existing EU copyright legislation. For example, Article 5(3)(a) of the InfoSoc Directive that allows Member States to provide for an exception for the sole purpose of illustration for teaching or scientific research employs similar wording.\(^\text{106}\) However, as pointed out by Maurizio Borghi and Stavroula Karapapa, the way data analysis works frequently makes acknowledging of the source practically impossible.\(^\text{107}\) Hence, if formulated in a restrictive manner, the attribution requirement may pose an insurmountable obstacle to data analysis. For instance, the Database Directive takes a hard stand on the attribution requirement by omitting the phrase ‘unless this turns out to be impossible’ and considers the requirement of source indication

\(^{105}\text{Fundamental research, also called pure research or basic research, is scientific research aimed to improve scientific theories for improved understanding or prediction of natural or other phenomena.}\)

\(^{106}\text{Article 5(3)(a) of the InfoSoc Directive reads as ‘use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible and to the extent justified by the non-commercial purpose to be achieved’}\)

\(^{107}\text{Ibid. p 61}\)
having an absolute character. Fortunately for data analysis in the UK, Article 29A(1)(b) follows the approach established in Article 5(3)(a) of InfoSoc Directive and mitigates the requirement to acknowledge the source and author’s name by exempting the cases where ‘this would be impossible for reasons of practicality’.

Finally, the UK TDM exception explicitly restrains contractual freedom of parties. Contract law allows parties to deviate significantly from provisions of the applicable legal framework. Intellectual property law may be supplemented or on the contrary derogate by contractual clauses. Using its contractual freedom, Europe PubMed Central (Europe PMC) explicitly excludes data analysis. According to publisher’s contract, ‘[c]rawlers and other automated processes may NOT be used to systematically retrieve batches of articles from the Europe PMC web site. Bulk downloading of articles from the main Europe PMC site, in any way, is prohibited because of copyright restrictions’. In contrast, STM, a big publisher, permits data analysis by stipulating in its Statement Sample License that ‘[s]ubscriber may access and use the Publisher’s content for TDM provided hereunder in accordance with the provisions and during the term of this Agreement for as long as the Subscriber contemporaneously maintains a subscription to such Publisher Content on the "XYZ" online service. Hence, contract law has been used to both restrict and permit data analysis. However, contractual law is mostly viewed as an obstacle for data analysis rather than an enabler. As pointed out by Estelle Derclaye, even if a data analysis exception is implemented by a Member State and a researcher complies with all provisions, data analysis may still be barred as any provision may be overridden by contractual agreement. Derclaye holds the opinion that the current legal framework makes it nearly impossible to perform data analysis with regard to sui generis rights. After considering briefly how contractual freedom affects data analysis, it appears that the UK government adopted a sound approach to the issue of data analysis by introducing a non-overridable copyright exception.

4.3 The EU

In September 2016, as a part of its update of EU copyright rules, the European Commission proposed a draft for a Directive on copyright in the Digital Single Market, which, inter alia, contains an exception that permits researchers to analyze on a large-scale scientific data to which they have lawful access. In Article 3, the Digital Single Market Directive stipulates that

---

108 Article 6(2)(b) of the Database Directive reads as ‘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’

109 Thorburn et al. (2017) p.132
“Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have lawful access for the purposes of scientific research. […] Any contractual provision contrary to the exception […] shall be unenforceable”.

The Digital Single Market Directive defines the phenomenon in Article 2(2) as

“‘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”

The definition differs advantageously from many definitions that are given in literature and legislation. I welcome the fact that it uses the term ‘automated’ because as explained earlier it makes the definition more technological neutral allowing it to include into its scope technologies that do not yet exist. The term ‘technique’ is, as explained earlier, narrow and does not reflect fully the complexity of the phenomenon. Using the term ‘processing’ would be preferable in the given context. Calling the phenomenon ‘text and data mining’ is not fully accurate because text is a subset of data and ‘mining’ emphasizes extraction process while an analytical analysis lays in the heart of any research. ‘Data analysis’ would be a more precise name for the phenomenon. The way the definition specifies sources of TDM could be improved. It would benefit the definition to name possible sources of TDM in an explicit and inclusive manner as for example – “video, texts, sounds, images or other elements, or a combination of these elements”.

Probably the most noticeable feature of the proposed exception that differs it from both Japanese and the UK exceptions is its scope of beneficiaries. While under the UK and Japanese approach, as long as other conditions are met, any individual can benefit from the respective exceptions, the proposed EU exception recognizes only research organisations, meaning that only individuals associated with a research organisation can benefit from the exception. Unlike the approach used in the research exception in the InfoSoc Directive, which defines beneficiaries implicitly by referring to noncommercial purpose of the activity, the approach taken in the proposed exception is straightforward. The proposal is so explicit regarding beneficiaries that it goes as far as to provide a definition of a research organisation. Article 2(1) reads as

“‘research organisation’ means a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to conduct scientific research and provide educational services:
(a) on a non-for-profit basis or by reinvesting all the profits in its scientific research; or
(b) pursuant to a public interest mission recognised by a Member State;

in such a way that the access to the results generated by the scientific research cannot be enjoyed on a preferential basis by an undertaking exercising a decisive influence upon such organisation”.

More explanation on research organisation is found in Recital 8 of the proposed Directive. The recital states that the proposed Directive acknowledges that a great diversity of research organisations exist in the EU. Different forms and structures of organisations may rise the question whether a certain organisation should be deemed as a ‘research organisation’ within the meaning of Article 3 of the Digital Single Market Directive. Recital 8 suggests two conditions that should be assessed when deciding on the status of an organisation. First, it has to be considered whether an organisation acts on nonprofit bases. Second, whether the organisation in question is carrying out any public-interest mission that is recognized by the State. The recital explains ‘public-interest mission’ by stating that factors such as public funding, provisions in national laws or public contract indicate such mission.

I welcome the explicit approach taken by the EU Commission in the proposed exception on TDM because it should make application of the proposed exception easier. For comparison, an application of the research exception in copyright requires considering the nature of the activity “as such”. Despite its well-defined scope regarding beneficiaries, fewer researchers can benefit from the EU proposed exception if compare to the UK or the Japanese exception. In case of the UK exception, any individual regardless whether he or she belong to a ‘research organisation’ can benefit from the exception, while the Japanese exception has even more favorable conditions for researchers because it does not have ‘non-commercial’ requirement. For me to decide whether such a narrow scope in relation to beneficiaries of the exception is appropriate and justifiable, I need to refer to Impact assessment “On the modernisation of EU copyright rules” (the Assessment) issued by the EU Commission to accompany the proposed Directive.

The Assessment identifies problems that are related to TDM and considers four different policy options that can solve these problems. First and second policy options are not relevant for identifying reasons why the proposed exception is limited to ‘research organisations’. To explain briefly what the two first policies are about, first policy option is concerned fostering industry self-regulation, the second policy option considers introducing a mandatory exception that would cover

---

text and data mining for non-commercial scientific research purposes. The third policy option includes limiting the scope of a future mandatory exception to ‘research organisation’ and was used as a foundation for Article 3 of the Digital Single Market Directive. The fourth policy option advocates a mandatory exception, which is applicable to anybody who has lawful access. Third and fourth policy options are scored approximately the same in the Assessment. The decisive factor for choosing the third policy option over the fourth policy option is that the third option is considered to have milder impact on the rightholders. According to the Assessment, a broad scope of the exception would have a significant negative impact on right holders because publishers would no longer be able to license TDM for scientific research purposes to commercial players, which represent an essential market for them, notably in areas such as life science and pharmaceutical.111

After the draft of the Digital Single Market Directive was published, several committees of European Parliament issued their draft opinions on the proposal. In its draft opinion, the Committee on the Internal Market and Consumer Protection argues that limiting the scope of Article 3 of the Digital Single Market Directive to research organisations is counterproductive.112 The committee suggests introducing a simple rule, which does not discriminate between users or purposes.113 Amendment 31 proposed by the committee excludes “made by research organisations” and “for the purposes of scientific research” and is formulated in the following way:

“Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC, Article 4(1) of Directive 2009/24/EC and Article 11(1) of this Directive for reproductions and extractions in order to carry out text and data mining of works or other subject-matter to which they have lawful access”.114

Another draft opinion was issued by the Committee on Culture and Education. This draft opinion generally agrees with the EU Commission on limiting beneficiaries of the new TDM exception to research organization as it decides to keep ‘research organisations’ as part of the wording of Article 3(1) of the Digital Single Market Directive. The Committee suggests changing the wording of Article 3(1) by adding the verb “acquired”. I my opinion, doing so the Committee tries to protect economic interests of copyright holder by indicating explicitly for access to be ‘lawful’ it has to

111 The EU Commission (2016) p. 106
113 Ibid.
114 Ibid. p. 25
has to be acquired, i.e. purchased. The amendment to Article 3(1) suggested by the Committee reads as

“Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject-matter to which they have acquired lawful access for the purposes of scientific research”.

Similar to the UK exemption, the EU proposal states explicitly that the rights granted under the exception cannot be contracted away. Considering non-overridable character of the EU exemption, it is appropriate to remind that contract law is a powerful legal regime that enables intellectual property owners to exclude the exercise of limitations on copyrights and sui generis database rights.

A group of intellectual property owners, namely publishers, is changing their business practices with the view that data analysis will become one of the main future vehicles for interaction with their intellectual property. It is reasonable to believe that publishers and other intellectual property owners will resort to contract law to safeguard their revenues from intellectual property when data analysis becomes a popular research technique.

In contrast to the UK’s focus on the individual, the EU proposed exception focuses on the research organisation. In order to understand content and logic behind employing the term ‘research organisation’ it is worth to consider the impact assessment on the modernisation of EU copyright rules, (the impact assessment) the document that accompanies the Digital Market Directive and explains policy rationale behind the draft legislation. According to the impact assessment, term ‘research organisation’ was adapted to replace non-commercial condition. Using only ‘research organisation’ was meant to include among exception’s beneficiaries Public-Private Partnership (PPP). Such approach should allow the exception to cover research projects, which may have an ultimate commercial outcome. While the UK approach is more favorable as it allows benefiting any individual who conducts data analysis irrespective of whether he or she belongs to any research institution, the EU approach has the potential to cover more instances of data analysis as carried out jointly by research institutions and private entities. The condition ‘non-commercial’ is not desirable in the new exception as it introduces similar obstacles that are associated with scientific research exception
5 Conclusions

Article 3 of the Digital Market Directive is certainly able to alleviate the legal uncertainty that surrounds data analysis. The proposal address adequately many legal issues that inhibit data analysis in Europe. The proposal includes an extended scope in relation to beneficiaries if compare to the copyright scientific research exception. Thus, the proposed exception covers data analysis that is carried out jointly by commercial and public entities. The proposal abandons traditional for copyright exceptions ‘non-commercial’ condition making it easier for judiciary to apply the exception because judges no longer have to consider the “nature of the activity as such”, but only need to look at the participants of a research project. Unfortunately, the proposal also has some drawbacks. While the exception is very explicit in relation to beneficiaries, it is not inclusive enough because it creates a privileged group of beneficiaries – research organisations. The reason behind limiting the exception to research organisations is to safeguard economic interests of copyright holders and database owners. However, it is difficult to understand why different entities that pay same amount of subscription fees receive different scope of right in exchange for their payments. Failing to include all the relevant beneficiaries the proposal stops short from unlocking innovation and merely promotes research.
References

Case law

Boosey v. Whight
Apple Computer, Inc. v. Computer Edge Pty., Ltd.
The UK High Court, 1900 (1899)
High Court of Australia, 6 May 1983, F.S.R. 537

Football Dataco Ltd and Others v Yahoo! UK Ltd and Others
The Court of Justice of the European Union, 1 March 2012, C-604/10

The Court of Justice of the European Union, 16 July 2009, C-5/08

Infopaq International A/S v. Danske Dagblades II (2012)
The Court of Justice of the European Union, 17 January 2012, C-302/10

Anne Frank Fonds v. Anne Frank Stichting and the Royal Netherlands Academy of Arts and Sciences
The district court of Amsterdam, 23 January 2015, C / 13/583257 / HA ZA 15-270

Legal instruments

TRIPS

The Berne Convention
The Berne Convention for the Protection of Literary and Artistic Works, Berne. 9 September 1886


The UK Copyright, Designs and Patents Act 1988, as amended in June 2014


Books


Chapters in books and articles


Jennings, Frank, John Yates. *Scraping over data: are the data scrapers’ days numbered?* In: Journal of Intellectual Property Law & Practice. Vol.4(2), pp. 120-129


Internet Sources

The International Federation of Reproduction Rights Organisations. *What is reprography?*. 2010. [http://www.ifrro.org/content/what-reprography](http://www.ifrro.org/content/what-reprography) [Visited 26 June 2017]


