Funded by

DE WOLF & PARTNERS

Study on the legal framework of text and data mining (TDM)

March 2014

Jean-Paul Triaille, partner, De Wolf & Partners, lecturer, University of Namur

with

Jérôme de Meeûs d'Argenteuil, associate, De Wolf & Partners

and with the collaboration of

Amélie de Francquen, associate, De Wolf & Partners, assistant, University of Namur

The information and views set out in this study are those of the authors and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use which may be made of the information contained therein

DOI : 10.2780/1475

ISBN : 978-92-79-31976-1

Table of Contents

I. TERMS OF REFERENCE OF THE STUDY

The background of this Study is described as follows in its Terms of Reference ("ToRs"):

> *"Data mining is currently subject to discussions in the UK in the context of a review of the copyright legislation in that Member State. Other Member States (*e.g. *Ireland) are also assessing the issue. The matter is nevertheless rather new and there is not a "coined definition" of what data mining activities are.*
>
> *Text and data mining has, in the impact assessment performed by the UK IPO, been defined in the following way:*
>
> *"Text and data and data analytics methods extract data from existing electronic information, to establish new facts and relationships, building new scientific findings from prior research. These new methods involve copying of prior works as part of the process to extract data".*"

The tasks to be completed, according to the ToRs of the Study, are the following:

> *1. Establish a working definition of "text and data mining" and, in this context, assess whether a distinction should be made between mining of text and mining of data.*
>
> *2. Assess whether mining of text and data freely accessible online (e.g., on twitter or facebook) should be distinguished from mining of text and data to which access is restricted (e.g., accessible only on the basis of a subscription).*
>
> *3. Assess what acts and corresponding rights could be relevant for text and data mining activities, e.g., the reproduction right provided for in Article 2 of the Infosoc Directive and in Article 3 a) of the Database Directive, as well as the sui generis right provided for in Article 8 of the Database Directive.*
>
> *4. Assess whether text and data mining activities (if determined under 3 that text and data mining required acts which are restricted by copyright) could be covered by the current exceptions and limitations to copyright and/or to the sui generis right. Examples of possible exceptions and limitations to copyright could be the exception for temporary acts of reproduction in Article 5.1, the exception for scientific research in Article 5.3 a) of the Infosoc Directive and the exception for scientific research in Article 6.2 b) of the Database Directive. Possible exceptions to the sui generis right could be the exception for scientific research in Article 9 b) of the Database Directive.*
>
> *5. Analyze whether, and if so in what way, existing rules on copyright and databases' sui generis right could constitute an impediment to text and data mining activities.*
>
> *6. Assess how the legal positions of the different stakeholders are affected by the current situation.*
>
> *7. Verify whether there are explicit legal provisions (e.g. different from general provisions implementing the Information Society Directive or the Directive on the legal protection of databases such as generally worded exceptions "for research purposes") in the Member States, decisions or judgments affecting text and data mining. Please limit the examination to the following Member States: Germany, France, the UK, Italy, Spain, Poland, Denmark, Hungary and the Benelux.*
>
> *8. Based on the above, assess the need, and if so the possible options, for legislative changes including whether there is a need to establish a specific exception to copyright for text and data mining activities (and, if relevant, the sui generis right for databases). In this context, list a range*

of policy alternatives and how to ensure a balance of the different interests involved. Please also identify, when relevant, possible conditions in terms of e.g. non-commercial/commercial research, whether the research results obtained from text and data mining could substitute the text and data being mined, whether there is a need to clarify that there should be lawful access to the text/data being mined, whether there is a need for (technical) control of access by the rightholder etc., and

9. Analyse the relation between a possible exception for text and data mining and existing licensing practices from rightholders such as publishers or scientific journals, holders of commercial databases, etc.

Tasks 1 and 2 shall be carried out under Part II ("Definitions and access to data"), task 3 under Part III ("Acts and exclusive rights"). Task 4 shall be carried out under IV ("Exceptions and limitations"). Tasks 5 and 6 shall be carried out under Part V. Task 7 shall be carried out under Part VI on the other legal provisions relevant to data analysis. Parts VII, VIII and IX describe which initiatives (legislative or not) the European Commission could take, and our proposal for a new specific exception for text and data mining.

Where appropriate, we do refer to national implementation of the relevant EU Directives in a number of Member States, i.e. (as we were asked) Germany, France, the UK, Italy, Spain, Poland, Denmark, Hungary and the Benelux countries.

The research for this Study has been conducted until February 2014.

II. DEFINITIONS AND ACCESS TO DATA

A. TOWARDS A DEFINITION OF DATA ANALYSIS

In this first Part, we will endeavour to establish a working definition of "text and data mining" and, in this context, assess whether a distinction should be made between mining of text and mining of data.

For the reasons explained *infra*, we will refer in this Study to "data analysis" and not to "text and data mining". It is indeed our understanding that "text mining", "data mining" and "text and data mining" are subsets of a more general concept which is often referred to as "data analysis".

a) Why referring to "data" rather than to "text"?

Before we come to a definition of text and data mining, the Terms of Reference require that we examine whether a difference should be made between "mining of text" and "mining of data".

We think it would be restrictive to refer to "text" analysis, because it gives the false impression that analysis activities are exclusively applied to "text", whereas the analysis of not only texts but also of images, videos, photos, etc., through the use of existing or future technologies, should also fall within the scope of the definition. Although the objects are different, the analysis of texts, videos, images and pictures present basically the same legal issues in terms of copyright and database protection.

Moreover, the EU lawmaker has given a broad definition of "data" in the Directive of 11 March 1996 on the legal protection of databases[1] (the "Database Directive")[2]; a *data*base is by definition composed of "data", but in reality and in accordance with the Database Directive, it can be composed of elements such as texts, images, sounds, data, etc. In our view, "data" can therefore be considered a generic or general term and includes all types of contents such as text, images, video, etc.

Article 1.2 of the Database Directive defines "database" as:

> *"a collection of independent* ***works***, ***data or other materials*** *arranged in a systematic or methodical way and individually accessible by electronic or other means"* (emphasis added).

Recital 17 of the Database Directive further states that:

> *"(…) the term 'database` should be understood to include literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data".*

Jean-Paul TRIAILLE and Alain STROWEL commented that:

> *"The « works » can be musical, literary, artistic or other types of works and the database can also be composed of materials such as texts, sounds, images, numbers, facts and data (cf. Recital 17 of the Database Directive). The Commission has, on multiple occasions, explained that "materials" should be understood in its broadest sense"* (our translation)[3].

[1] Directive 96/9/EC of the European Parliament and of the Council of 11 march 1996 on the legal protection of databases, OJ L 077, 27/03/1996 p. 0020 – 0028.

[2] See also *infra* in this Study.

[3] A. STROWEL & J.P. TRIAILLE, *Le droit d'auteur, du logiciel au multimédia : droit belge, droit européen, droit comparé*, Cahiers du Centre de Recherches Informatique et Droit (CRID), Bruylant, Bruxelles, 1997, p. 260 : « Les 'œuvres' peuvent être des œuvres musicales, littéraires, artistiques ou autres et la base de données peut aussi être composée de « matières » telles que textes, sons,

Another commentator, Benoît MICHAUX, further explained that:

> *"Database means a collection, this seems to go without saying. This collection covers works and "non-works", designated under the Directive as "data or other materials" […] "the sort of elements integrated in the collection does not matter: it can include, e.g, texts, images, sounds, music or graphic designs"* (our translation)[4].

We can logically conclude that, for the purpose of this Study, it is better to talk about "data mining" than about "text mining" or than about "text and data" mining.

"Text mining" is in turn too specific as it sounds to exclude images, sounds, pictures, graphics, maps, and any other works or data (protected by copyright or not) than texts.

For the same reasons, and to answer a question raised by the Terms of Reference, "mining of text" is also too specific, compared to "mining of data".

Should one make a difference between these two terms? Yes, because, as was just explained, "mining of text" is just one subset of "mining of data". In terms of copyright, as will be further examined below, texts can be protected by copyright[5] or not, and data can be protected (by the database maker's right or *sui generis* right) or not. But neither text and data mining nor data mining, text mining or the scope of the Study are limited to elements which are protected by an intellectual property rights, so that the protected/unprotected status is irrelevant.

b) Why referring to data "analysis" rather than to data "mining"?

In our view, and for the sake of legal security and "future-proofness", the definition of text and data mining should be technology-neutral, evolutive and made for changing technologies. As explained by Maurizio BORGHI and Stavroula KARAPAPA:

> *The 'challenge of better lawmaking' is that of enacting provisions at a sufficient level of generality to make copyright principles applicable to unforeseen situations".*[6]

We therefore suggest referring to "data analysis" rather than to "data mining". Indeed, it is generally admitted that "to mine" means "to extract data from texts *qua* informational resources"[7], whereas data *analysis* is encompassing much more than the mere extraction of data; it may vary from one technology to another, and may further extend to other activities than "mining" or "extracting", depending on the context, the technology and the sectors involved (see some of the definitions quoted *infra*, which make reference to extract, but also to crawl, process, compare, copy, analyze, retrieve, interpret, search, sort, parse, remove, etc.).

---

images, chiffres, faits et données (cf. considérant 17 DBD). La Commission a, à diverses reprises, expliqué que le terme « matière » devait être compris au sens le plus large ».

[4] B. MICHAUX, *Droit des bases de données*, Kluwer, 2005, p.3 « La base de données consiste dans un recueil, cela semble aller de soi. Ce recueil intègre des œuvres et des « non-œuvres », désignées par la directive sous l'expression « données ou autres éléments ». […] Le genre des éléments intégrés dans le recueil importe peu : il peut s'agir, par exemple, de textes, d'images, de sons, de musiques ou de graphismes ».

[5] The main categories of texts which will not be protected by copyright are : (1) texts which belong to the public domain (i.e. whose author died more than 70 years ago); (2) (although this is not always the case, and notably not in the UK) official texts in the sense of article 2.4. of the Berne Convention (which states as follows: "*It shall be a matter for legislation in the countries of the Union to determine the protection to be granted to official texts of a legislative, administrative and legal nature, and to official translations of such texts".* One could add that very short texts, composed of a few words might often not be protected; but then, should one talk in such case about texts or rather about "sentences" or "short sentences"?

[6] BORGHI, M., and KARAPAPA, S., *Copyright and Mass Digitization : a Cross-Jurisdictional Perspective*, Oxford University Press, 2013, p. 63.

[7] BORGHI, M., and KARAPAPA, S, op. cit., note 6, p. 47.

Furthermore, it is often understood that "to mine" content means "to go deep into" texts, video, images, photos, etc. whereas some data analysis techniques "stay on the surface" of texts, video, images, photos, etc.[8]

For the purpose of this Study, we will hereafter often refer to "data analysis" rather than to "data mining" or to "text and data mining" (TDM). Because the expression TDM is still often used as an acronym, we may sometimes continue to use it in this Study. Our suggestion is that "TDM" (or "text and data mining") should not be used in a legislative provision and that "data analysis" is a better expression for legislation. In legal doctrine, whatever terms are used does not really matter of course.

Our suggestion to refer to "data analysis" in a possible future legislation is comforted by some legislation or draft legislation. Recently indeed, a few European and non-European lawmakers have been investigating the issue of "text and data mining", mainly to introduce (or to envisage to do so) an exception to copyright for such purpose. Specific exceptions to copyright laws are presently under scrutiny both in Ireland and in the UK. To a lesser extent, Australia is also investigating the amendment of its copyright law regarding data analysis[9].

Japan is the only country where an exception for data analysis is in force (since 2011).

In their current versions, the texts of these bills or laws do not share a common definition of data analysis.

Before proposing a working definition of data analysis, we will give an overview of the way data analysis has been defined so far by the legislation in Japan, the UK and Ireland, and by the legal and scientific literature. We will try to highlight what terms we should include/exclude in the working definition and the reasons associated thereto.

European and non-European authors and researchers do not agree on the definition of data analysis (see *infra*). We will also list a number of these definitions before suggesting one.

c) Definitions in the legislation of Japan and draft legislation in the UK and Ireland

In the following paragraphs, we will analyze the definitions of data analysis given by the Japanese, the UK and the Irish lawmakers. We will see that the technology (by computer or by other means) and the purpose (commercial or non-commercial use) associated with data analysis techniques have prompted different solutions in the countries under consideration.

(i) 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":

> *For the purpose of information analysis ('information analysis' means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other 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*

[8] It is our understanding that data analysis techniques can e.g. include crawling techniques which do not *per se* always involve much more than "surfing on the surface of texts".

[9] Australian Law Reform Commission (ALRC), http://www.alrc.gov.au/publications/8-non-consumptive-use/text-and-data-mining

a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary(…) "[10].

The *rationale* for this exception was that:

"copyright shall not be exercised against the act of reproducing or adapting a work on a recording medium for the purpose of information analysis to be conducted on a computer under certain conditions"[11].

Article 47-7 ends with the following sentence*:*

"However, an exception is made of database works which are made for the use by a person who makes an information analysis"

The meaning of this sentence is not very clear and happens to be translated or explained in various ways[12]. Our understanding of the text is that the exception (allowing reproduction for data analysis purposes) applies to all works except to databases which are precisely made to be used for data analysis purposes.

This provision is interesting: it is, in our opinion, based on a similar logic than when e.g. a legal provision introduces an exception allowing reproduction of books for educational purposes but discards the exception for scholarly books. The market and the *raison d'être* of scholarly books are the educational world; therefore, allowing to copy school books for educational purposes would completely undermine the market and certainly conflict with the normal exploitation of such works and unreasonably prejudice the legitimate interests of the author (article 9.2. Berne Convention). In the same way, the Japanese TDM exception discards its application to databases which are precisely made to be used for data analysis. We will come back on this discussion when formulating recommendations (see Part IX).

The approach taken by the Japanese lawmaker is both extensive and limitative in the way it delineates the concept of data analysis:

- *Extensive* because it gives a broad definition of "information" (referring to languages, sounds, images or other elements constituting such information) and "analysis" (covering the act of extraction, comparison, classification or other statistical acts of analysis). Moreover, it expressly authorizes the reproduction ("to make recording on a memory") and the adaptation of the work ("including a recording of a derivative work created by such adaptation") for the purposes of data analysis – this will be further examined in Part IV of this Study.
- But also *limitative* because the "reproduction, etc., for information analysis" must be made "by using a computer". As we shall see, a better proposal comes from the UK lawmaker who simply refers to "automated analytical techniques" or "electronic analysis"; which is less technology-dependent and gives more legal security (see *infra*).

As explained by Maurizio BORGHI and Stavroula KARAPAPA:

[10] See Copyright Research and Information Center, "Copyright Law of Japan", Article 47septies, http://www.cric.or.jp/english/clj/cl2.html#cl2_1+A47septies.

[11] AIPPI Japanese Group "The types of works subject to this provision are not limited. The Subdivision on Copyright of the Council for Cultural Affairs issued a Study in January 2009 where it presents the following examples of information analysis: (1) website information analysis and language analysis in which the use of a specific language or character string is analysed and statistically processed and (2) sound analysis and video/image analysis in which the meaning of the sound wave, video, character string, etc., comprising a certain sound, video, image, etc., is analysed. However, the types of information analysis should not be limited to these examples", in "Exceptions to Copyright protection and the permitted Uses of Copyright works in the hi-tech and digital sectors" (Question Q216B), 16 May 2011, p.9, https://www.aippi.org/download/commitees/216B/GR216Bjapan.pdf.

[12] "*However, an exception is made of database works which are made for the use by a person who makes an information analysis*."; or "*However, the work of a database that is widely available in information analysis, this limitation provision does not apply*. "; or "*But, about the book of a database offered for information analysis widely, this limit rule isn't applied*." ; or "*However, these restrictive regulations are not applied to the work of the database with which information analysis is provided widely*. "

> *"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."*[13]

The Japanese lawmaker seems to exclude raw data (i.e. data which are not protected by copyright) from the scope of the exception for data analysis since Article 47-7 of the Japan Copyright Act refers to the recording and the adaptation of a "work" – which is defined in Article 10 of the Japan Copyright Act :

> *"(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.*
>
> *(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*".[14]

However, the limitation of the exception to protected works can in our view be explained by the fact that there is no database protection legislation in Japan similar to the *sui generis* protection of the Database Directive, so that only copyright comes into play (and requires an exception if the purpose is to exempt data analysis from the copyright exclusive rights). Let us mention that the proviso that data analysis should be made possible "to the extent deemed necessary" may run contrary to the objective of legal security and lead to interpretation issues.

(ii) The UK

In 2010, the UK Prime Minister commissioned a report to consider whether the UK's intellectual property framework was up to the task of supporting innovation and growth (the "Hargreaves report"). One of the findings of the Hargreaves report (published in May 2011) was that the UK Government should introduce an exception to allow the use of analytics for non-commercial use. Since then, the UK IPO has made several attempts to define the scope and the limits of this exception.

In August 2011, the UK Government's response to the Hargreaves Report pointed out that:

> *"Automated 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"*[15]*.*

We see here that the scope of the definition refers to "text and data mining" as one example of "automated analytical techniques". However, the reference to "automated analytical techniques" covers a broader range of technologies than just "computer" (cf. Japan definition) – this wording can thus be more appropriate for a working definition. The definition gives as examples of what can be analyzed "articles in scientific journals and other works (i.e. exclusively literary works), but these are only mentioned as examples of "electronic information" which is a much more encompassing terminology.

Contrary to what is stated in the UK Government's response to the Hargreaves Report, we do not think that data analysis always involves the *copying* of (electronic information). We will come back on this subject later in this Study (Part III "Acts and corresponding rights").

---

[13] BORGHI, M., and KARAPAPA, S, op. cit., note 6, p. 62.

[14] See Copyright Research and Information Center, "Copyright Law of Japan", Article 47septies, http://www.cric.or.jp/english/clj/cl2.html#cl2_1+A47septies. Paragraph 2 obviously comes from article 2.8 of the Berne Convention which states that "(t)he protection of this Convention shall not apply to news of the day or to miscellaneous facts having the character of mere items of press information".

[15] Royal Coat of Arms of the United Kingdom (HM Government), "Modernising Copyright: A modern, robust and flexible framework", Government response to consultation on copyright exceptions, Annex E: "Data analytics for non-commercial research", August 2011, p. 36.

In December 2012, the UK IPO made an impact assessment for the introduction of the "Exception for copying of works for use by text and data analytics". For that purpose, it defined "text and data and data analytics methods" as methods to:

> *"extract data from existing electronic information, to establish new facts and relationships, building new scientific findings from prior research. These new methods involve copying of prior works as part of the process to extract data*".

For the reasons stated above, we do not think that data analysis systematically involves the act of extraction. The reference to "extract", except when used as an example of techniques possibly underlying data analysis methods, should thus be avoided accordingly. Similar to what we said above, we think that data analysis does not always involve the copying of work. Introducing such element in the definition partly preempts the consequences, may not always be accurate and risks also being outdated in the future as technologies evolve.

In June 2013, the UK IPO published, for technical review, a draft exception which amends the Copyright, Designs and Patents Act by adding a new Section 29A and a new Annex 2C on "Data analysis for non-commercial research"[16]:

Section 29A (Data analysis for non-commercial purposes) of the draft provides:

> *(1) Where a person has lawful access to a copy of a copyright work, copyright is not infringed where that person makes a copy of the work for the purposes of carrying out an electronic analysis of anything recorded in the work provided that:*
>
> *(a) it is done for the sole purpose of non-commercial research; and*
>
> *(b) the copy is accompanied by sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise).*
>
> *(2) Any dealing with a copy made pursuant to section (1) for a purpose other than the purpose referred to in subsection (1) is an infringement of copyright and where such a copy is permanently transferred to another the copy shall be treated as an infringing copy.*
>
> *(3) To the extent that the term of a contract purports to restrict or prevent the doing of any act which would otherwise be permitted by this section, that term is unenforceable"*

Annex 2C of the draft provides:

> *(1) Anything which, by virtue of section 29A (data analysis for non commercial research), may be done in relation to a copyright work without infringing copyright in that work, may be done in relation to a work in which rights are conferred by this Chapter without infringing those rights.*
>
> *(2) Where by virtue of section 29A a copy made pursuant to that section is to be treated as an infringing copy, such a copy shall be treated as an illicit copy for the purposes of this Chapter.*
>
> *(3) To the extent that the term of any contract purports to restrict or prevent the doing of an act which would otherwise be permitted by this paragraph, that term is unenforceable.*

The copying is only allowed, under this exception, when the person "has lawful access to a copy of a copyright work" (see *infra,* Part IV). On the other hand, it refers to "electronic analysis" which is broader than just analysis made "by computer" (as in Japan). And the terms "anything recorded in the work" are broad enough to include texts, videos, images, etc. The scope of this exception is further restricted by the fact that "data analysis" can only be done for "the purpose of non-commercial research" and to the extent that "the copy is accompanied by sufficient acknowledgement"; we will come back to this later as these two criteria are not relevant as such for the purpose of a definition of data analysis and will be commented in Part IX of our Study. Finally, this exception restricts the contractual freedom of the parties to the extent that it provides that "*a term of a contract [which] purports to restrict or prevent the doing of*

[16] UK IPO website, http://www.ipo.gov.uk/types/hargreaves.htm

any act which would otherwise be permitted by this section […] is unenforceable"; we shall also analyze this unwaivability issue in Part IX.

For terminology purposes, it is worth noting that the texts have referred to "automated analytical techniques", "text and data and data analytics methods", "data analysis" (for non-commercial research) and to "electronic analysis", etc.

(iii) Ireland

In 2012, the Irish Copyright Review Committee issued a Consultation Paper on "Copyright and Innovation"[17] for the Department of Jobs, Enterprise and Innovation containing a set of recommendations, including on the subject of "Text and Data Mining". The text of the Consultation Paper proposes to amend the Irish Copyright and Related Rights Act (CRRA) by adding a new section 106F on "Digital analysis and research".[18]

Section 106F (entitled "Digital analysis and research") would provide that:

> *(1) It is not an infringement of the rights conferred by this Act for a person to reproduce a work for a purpose to which this section applies if:*
>
> *(a) it would not be practical to carry out the research without making the copy,*
>
> *(b) the person is the owner or lawful user of the work, and*
>
> *(c) the person has informed the owner of the rights in the work, unless this is unreasonable or inappropriate or turns out to be impossible for reasons of practicality or otherwise.*
>
> *(2) This section applies to:*
>
> *(a) text-mining, data-mining, and similar analysis or research,*
>
> *(b) encryption research and similar analysis or research, and*
>
> *(c) such other analysis or research as the Minister may by order provide.*
>
> *(3) Nothing in Part VII shall be construed as operating to prevent any person from undertaking the acts permitted by this section or from undertaking any act of circumvention required to effect such permitted acts"*

As explained *supra*, we take the view that data analysis does not always involve an act of reproduction. Regarding definitions, the references to "text-mining, data-mining, and similar analysis or research" do allow a wide interpretation and may be considered as being technology neutral. The reference to "encryption research and similar analysis or research" is alien to data analysis (as we understand it in this Study); the reference to "such other analysis or research as the Minister may by order provide" could be a manner to, in the future, adapt data analysis (as we understand it here) to yet other technologies or practices[19].

It is interesting to note that, in these three legislative texts (one statute, two draft bills), none refers to the commonly used terminology of "text and data mining" but instead refers to more encompassing terms. This is important for proposing a definition. In our view indeed, a "working definition" should be a definition which can be used in a legislative text (be it a national legislation, an EU directive or any other legislation). It is certainly no coincidence that no legislator used the word "text and data mining" as cover word to describe what they wanted to regulate.

[17] Copyright Review Committee, "Copyright and Innovation: a Consultation Paper", for the Minister for Jobs, Enterprise and Innovation, Dublin 2012.

[18] http://www.djei.ie/science/ipr/crc_statement.htm

[19] In our view, the possibility given to the Government (here, a Minister) to widen the scope of a statute raises however questions and should generally be avoided (for it creates legal uncertainty and may lack parliamentary legitimacy).

d) Some definitions in the legal and scientific literature

Many authors, researchers, legal commentators and scientists have tried to define the concept of data analysis. We will hereafter list a number of definitions which we thought could be of interest for trying to come to "a working definition of text and data mining", as asked for by the Terms of Reference of the Study:

> "*Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources. A key element is the linking together of the extracted information together to form new facts or new hypotheses to be explored further by more conventional means of experimentation*". **Marti A. Hearst.**[20]

This definition is limitative as it refers to "by computer", "extracting", and "written resources" – see comment *supra*. Moreover, the reference to "more conventional means of experimentation" is vague and does not pertain to a definition (such exploration might not always take place).

> "*Text mining attempts to discover new, previously unknown information by applying techniques from information retrieval, natural language processing and data mining*". **UK National Center for Text Mining (NaCTeM)** [21]

This definition is text-specific ("text-mining") and refers to a non-exhaustive list of techniques "from information retrieval, natural language processing and data mining". Data mining is here only a subset of text mining (while we would tend to consider that the opposite is true instead).

> *"A computational process whereby text or datasets are crawled by software that recognizes entities, relationships and actions*". **The International Association of Scientific, Technical and Medical Publishers (STM)** [22]

This definition confirms our assumption that "data analysis" does not always "go deep into" content as it refers to "crawling" and not necessarily "extracting"[23].

> *"Text and Data Mining means to perform extensive automated searches of Publisher's Content, the sorting, parsing, addition or removal of linguistic structures, and the selection and inclusion of content into an index or database for purposes of classification or recognition of relations and associations*". **The International Association of Scientific, Technical and Medical Publishers (STM)[24]**

The definition comes from a sample standard text and data mining license drafted by the STM (International Association of Scientific, Technical & Medical Publishers). In this definition, "automated

[20] HEARST, M. A. , *What is Text Mining*?, SIMS, UC Berkeley, October 17, 2003.

[21] http://www.nactem.ac.uk/faq.php?faq=1

[22] International Association of Scientific, Technical and Medical Publishers (STM), "Submission on the Issues Paper "Copyright and the Digital Economy" (UK)", Oxford, 29 November 2012.

[23] Wikipedia defines a Web crawler as "an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web indexing" (a "bot" is here equivalent for a "robot"). The definition in Wikipedia adds that "Web crawlers can copy all the pages they visit for later processing by a search engine that indexes the downloaded pages so that users can search them much more quickly". Consequently, we think that the act of "crawling" stays on the surface of the content, whereas the act of "extraction" goes deep into the content.

[24] International Association of Scientific, Technical & Medical Publishers (STM), "Text and Data Mining Sample Subscription", 15 March 2012, http://www.stm-assoc.org/2012_03_15_Sample_License_Text_Data_Mining.pdf.

search" fits better than "by computer" (see Japanese law) as it does not limit the activity to computer technologies. The definition covers all sorts of works or data (see definition of "Publisher's Content" in the license) and refers to a non-exhaustive list of techniques ("search, sorting, parsing, addition or removal of linguistic structures, and the selection and inclusion of content into an index or database") and purposes ("for purposes of classification or recognition of relations and associations").

> "*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*". **Jonathan Clark**[25]

Interestingly, this author makes a different distinction between "text mining" and "data mining".

> "*Automated tools, techniques or technology to process large volumes of digital content that is often unstructured or not uniformly structured with one or more of the following purposes: (i) to identify and select relevant information, (ii) to extract information from the content, and (iii) to identify relationships within/between/across documents and between incidents or events for meta-analysis*". **Eefke Smit and Maurits Van der Graaf** [26]

For the references to "automated" and to "digital content", see our comments *supra.* These authors make an interesting distinction between "tools, techniques or technology". However, the term "technology" is probably a generic term that may be seen to encompass the two other terms. The acts of identification, selection and extraction might not be the only purposes of data analysis.

> "*[A]utomated processing of large amounts of structured digital textual content, for purposes of information retrieval, extraction, interpretation, and analysis*". **Bernard F. Reilly** [27]

About "automated processing", see our comment *supra.* "Structured digital textual content" is in our view too limitative, both because non-structured content could be processed and analyzed and because non-textual content may also be processed and analyzed. And "purposes of information retrieval, extraction, interpretation, and analysis" might not be sufficiently exhaustive and exclude activities which should normally fall under the concept of data analysis.

> "*[T]he extraction of data from large datasets to uncover previously unknown and potentially useful information*". **Andres Guadamuz and Diane Cabell** [28]

"Extraction" is in our view too limitative – see comment *supra*. "Data" refers to all types of content, which is a positive element – see our comment *supra.*

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

[26] SMIT, E. and VAN DER GRAAF, M., "Journal article mining: the scholarly publishers' perspective", Learned Publishing vol . 25 no. 1, January 2012, p.36.

[27] REILLY, B. F., "When Machines do Research, Part 2: Text-Mining and Libraries," Charleston Advisor, October 2012, pp. 75-76.

[28] The authors identifies various types of content that are subject to automated analysis: "(i) *Text*: published articles, book chapters, preparatory notes, working papers, reports, teaching materials, conference papers, presentations, theses; (ii) *Datasets*: Statistical data, geolocation data, survey results, maps, figures, time series, genetic information, health records, computer logs; and (iii) *Multimedia*: pictures, sound recordings, interviews, presentations, video". They add that "Each of the above may have separate legal regimes applying to them. In the interest of convenience and simplicity, whenever the report talks about database contents, there will be no distinction as to whether we are dealing with text, data or multimedia, unless clearly specified in the text". GUADAMUZ A., and CABELL, D. "Analysis Of UK/EU Law On Data Mining In Higher Education Institutions" (WHITE PAPER), January 15, 2013, p.4.

> "*Text-mining, data mining or media mining is the extracting of "chunks" of data using computer programmes to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modelling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow facts or hypotheses to be discovered or* analysed". **British Library**[29]

Here, the British Library makes a distinction between *text* mining, *data* mining and *media* mining and defines it as the "extraction of data". This confirms our previous assumption that text mining, data mining and media mining are subsets of data analysis. There are limitations to the effectiveness of the definition due to the use of the words "computer programmes". Furthermore, the definition is limitative in the sense that it only seems to refer to "extraction" - see comment *supra*. The list of technological means ("machine learning, statistical analysis, modeling techniques and database technology") is descriptive, might not be exhaustive nor as such necessary in a definition but it does help in clarifying what the notion covers.

e) Coming to a definition of "text and data mining"

In view of the foregoing, we propose the following definition for "data analysis" (which we use as a terminology instead of "text and data mining"):

> "*The automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights*."

The elements of this definition shall be interpreted as follows:

(i) *automated:*

data analysis is done by applying automated techniques to a set of selected digital materials. Computers are one of the automated techniques available, but there may (and in the future probably there will be) other techniques; "automated" is opposed to "made by humans" and this is indeed this characteristic which makes TDM so powerful and which raises new IP (intellectual property) issues;

(ii) *processing:*

data analysis involves the processing of data, which may (but this will not always in every case) include the extraction, copy, comparison, classification or other statistical analysis, etc. of data, or a mix of them;

(iii) *of digital materials, which may include texts, sounds, data, images or other elements:*

it means that data analysis can be applied to all types of contents; we use the plural but data analysis can also be performed on the basis of one document (one book, for example); in most cases, the interest of the process is however to include a large number of materials;

(iv) or a combination of these:

this is often the case. Data analysis can be applied simultaneously on all types of contents.

(v) in order to uncover new knowledge or insights:

[29] British Library Submission to Hargreaves Report.

such as new entities, new relationships or new actions. No one does data mining for the sake of it, one does it to discover (uncover) new knowledge, new insights, new relationships, etc.

We deliberately avoided referring to the fact that such operations do (or often do) involve copyright materials, as well as references that they might lead to reproduction or creating derivative works, as these issues do not pertain to a definition (which, in our view, should avoid having built-in in legal terminology).

B. ACCESS TO DATA FOR DATA ANALYSIS PURPOSES

In this chapter, as requested by the Terms of Reference, we will "assess whether mining of text and data freely accessible online (e.g., on Twitter or Facebook) should be distinguished from mining of text and data to which access is restricted (e.g., accessible only on the basis of a subscription)".

The claim currently being made for data analysis (applied to texts) by the scientific community is that data analysis can speed up the research process and capitalise on work which has been done in the past in a new and effective way. However, a number of features need to be in place before this can happen.

These include:

> (i) *access to a vast corpus of research information, (ii) in a consistent and interoperable form, (iii) freely accessible, without prohibitive authentication controls, (iv) covering digitised text, data and other media sources, (v) unprotected by copyright controls (over creation of derivative works), (vi) a single point of entry with a powerful and generic search engine, (vii) a sophisticated mechanism for enabling the machine (computer) to analyse the collection for hidden relationships*". **ICSTI**[30]

We do not at this stage endorse the statement that the corpus should be freely accessible and unprotected by copyright controls, which in any case does not correspond to today's reality but the chronology described in the paragraph above seems clear and useful for the reader.

In the online environment, the access to data (texts, video, images, tables, etc.) can take multiple forms.

The Terms of Reference were asking us to distinguish between just two levels of access, i.e. data freely accessible online on one hand, and mining of text and data to which access is restricted, on the other hand. We have considered that it would be logical to distinguish not just two but four levels of access to data.

a) Four different levels of access

These four levels of access could be qualified as follows (the terminology is ours but helps in presenting the different situations): "all to all" / "many to many" / "one to many" / "one to one".

(i) "**All to all**": *data freely accessible on the Web* (we will call them "**web data**") can be accessed, practically, on a "no condition basis" (i.e. there are no "terms and conditions" attached). Anybody can access it.

(ii) "**Many to many":** *data created and shared on social networks* (we will call them "**social networks data**") is dependent on the private account settings of the users: public[31]/private[32]/custom[33] and the terms of use of the social networking platforms - such as

[30] International Council for Scientific Information (ICSTI), "Text and Data Mining", July 2009, p.1.

[31] Facebook Data Use Policy (Sharing and finding you on Facebook - https://www.facebook.com/about/privacy/your-info-on-fb) : "Choose this icon if you want to make something Public. Choosing to make something public is exactly what it sounds like. It means that anyone, including people off of Facebook, will be able to see or access it. Learn more about public information ».

(but not limited to) Facebook[34], Twitter[35], and Pinterest[36]. Alternatively, data mining users should be careful when "mining" data shared but not created on social networks – as they could be qualified as contract data and/or confidential data (*infra)*. Subject to the users' account settings, many or all people can see it.

(iii) "**One to many**": *contractual clauses* can restrict or limit the access to data and for instance prohibit data analysis or other text or data mining ("**contractual/publishers data**"). This may be the case for data covered by contracts with publishers and/or repositories. Only the authorised users can access the text, and prior to that, they will have accepted terms of use (at least in the on-line environment or for digital products).

(iv) "**One to one**": *confidentiality agreement and/or clauses* are used when a person (or a company) who is the only one to be in possession of some "**confidential data**", may decide to disclose it to another person or company. In such cases, parties will use "non disclosure agreements" (often called NDAs) and such NDAs will contain many restrictions regarding the use which can be made of the data being disclosed. Typically, they may not be disclosed to others, must remain technically protected, may only be used for specifically described purposes, may not be disassembled, etc.

We thought it was useful to enlarge the number of categories from two to four.

So, we have the following summarizing table:

[32] Facebook Data Use Policy (Sharing and finding you on Facebook - https://www.facebook.com/about/privacy/your-info-on-fb): "Choose this icon if you want to share with your Facebook Friends".

[33] Facebook Data Use Policy (Sharing and finding you on Facebook - https://www.facebook.com/about/privacy/your-info-on-fb): "Choose this icon if you want to Customize your audience. You can also use this to hide your story from specific people. If you tag someone, that person and their friends can see your story no matter what audience you selected. The same is true when you approve a tag someone else adds to your story".

[34] Facebook Statement of Rights and Responsibilities (available at: https://www.facebook.com/legal/terms) : "You own all of the content and information you post on Facebook, and you can control how it is shared through your privacy and application settings. In addition: For content that is covered by intellectual property rights, like photos and videos (IP content), you specifically give us the following permission, subject to your privacy and application settings: you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook (IP License). This IP License ends when you delete your IP content or your account unless your content has been shared with others, and they have not deleted it. […] When you publish content or information using the Public setting, it means that you are allowing everyone, including people off of Facebook, to access and use that information, and to associate it with you (i.e., your name and profile picture)".

[35] Twitter Terms of Service (available at: https://twitter.com/tos): "The Content you submit, post, or display will be able to be viewed by other users of the Services and through third party services and websites (go to the account settings page to control who sees your Content). You should only provide Content that you are comfortable sharing with others under these Terms. What you say on Twitter may be viewed all around the world instantly. You are what you Tweet!. […] You retain your rights to any Content you submit, post or display on or through the Services. By submitting, posting or displaying Content on or through the Services, you grant us a worldwide, non-exclusive, royalty-free license (with the right to sublicense) to use, copy, reproduce, process, adapt, modify, publish, transmit, display and distribute such Content in any and all media or distribution methods (now known or later developed). This license is you authorizing us to make your Tweets available to the rest of the world and to let others do the same".

[36] Pinterest Terms of Services (available at: http://about.pinterest.com/terms/): "Pinterest allows you to post content, including photos, comments, and other materials. Anything that you post or otherwise make available on our Products is referred to as "User Content." You retain all rights in, and are solely responsible for, the User Content you post to Pinterest. More simply put: If you post your content on Pinterest, it still belongs to you but we can show it to people and others can re-pin it. […] You grant Pinterest and its users a non-exclusive, royalty-free, transferable, sublicensable, worldwide license to use, store, display, reproduce, re-pin, modify, create derivative works, perform, and distribute your User Content on Pinterest solely for the purposes of operating, developing, providing, and using the Pinterest Products. Nothing in these Terms shall restrict other legal rights Pinterest may have to User Content, for example under other licenses. We reserve the right to remove or modify User Content for any reason, including User Content that we believe violates these Terms or our policies".

Access granted by [..] to [..]	***Types of data (our qualification)***
By all to all	Web data
By many to many	Social networks data
By one to many	Contractual/publishers data
By one to one	Confidential data

We will, to the extent such is relevant, reuse this distinction when dealing with the relation with a possible exception and the license agreements.

We will also describe (see point c. hereunder) two variations of these categories: public sector information on one hand, Creative Commons licences and Open Access licences on the other.

The level of access to data is paramount in order to determine the applicable legal framework (including, but not limited to intellectual property rules) and to assess the risk associated therewith. Indeed, general rules of intellectual property law may, in certain circumstances, be supplemented (and reinforced) or on the contrary derogated (alleviated) by contractual clauses.

Some contracts explicitly exclude data mining:

> "*Crawlers 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 web site, in any way, is prohibited because of copyright restrictions*". **Europe PubMed Central**[37]

On the contrary, other contracts specifically authorize data mining:

> "[*Licensee may] use the Licensed Material to perform and engage in text mining/data mining activities for academic research and other Educational Pur*poses". **NESLi2 License**[38]
>
> "*Subscriber 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*. **International Association of Scientific, Technical & Medical Publishers**[39]
>
> "*Archived content may not be published verbatim in whole or in part, whether or not this is done for Commercial Purposes, either in print or online. This restriction does not apply to reproducing normal quotations with an appropriate citation. In the case of text-mining, individual words, concepts and quotes up to 100 words per matching sentence may be reused, whereas longer paragraphs of text and images cannot (without specific permission from NPG*)". **Nature Publishing Group**[40]

Interestingly, the draft UK exception explicitly prohibits any contractual restrictions on the right of the users to mine data (subject to the conditions set out in the draft exception):

> "(3) *To the extent that the term of a contract purports to restrict or prevent the doing of any act which would otherwise be permitted by this section ["Data analysis for non-commercial research"], that term is unenforceable*".[41]

We will analyse this point later in this Study (Part IX).

---

[37] Europe PubMed Central labs, "Copyright", http://europepmc.org/Copyright

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

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

[40] Nature Publishing Group, "Terms for reuse of archived manuscripts", http://www.nature.com/authors/policies/license.html

[41] See http://www.ipo.gov.uk/types/hargreaves.htm

As explained above, contracts should be carefully scrutinized in order to determine the level of access to social data, contractual data and confidential data.

Before going further in this distinction, we shall clarify that we are not going to insert confidential data and non-disclosure agreements in our analysis. In these one-to-one negotiations, both parties have very specific constraints and purposes and it does not seem wise in a TDM context to intervene and regulate the scope of these agreements[42].

b) Can a distinction be made between the different levels?

As to whether a distinction can be made between text and data freely accessible online from mining of text to which access is restricted (e.g. only on the basis of a subscription), we think the answer is positive. Apart from the fact that, in practice, access is more or less easy, depending on the categories, also from a legal (and contractual) point of view the situation prevailing in the different categories varies:

Level 1 Web data: "All to All"

- Access to **web data** are generally not covered by contracts. As a matter of fact, irrespective of the existence of terms of uses governing the access to web data, we think it would be impossible to prove the consent of the users and the formation of the contract subsequently. The visitor of a website who was not asked to explicitly accept terms and conditions (available via a link on another page on the most visible pages of the site) nor to "click" in any way to show his consent could, in many jurisdictions, still argue that he is not bound by such terms and conditions, whatever they may be. It may be that there are variations in the laws of the Member States, as to when and to what extent the visitor to a website is bound by terms and conditions which are available on a page of the site but do not have to be accepted (by some "clickwrap" or other similar mechanism).

  Let us imagine that these terms and conditions prohibit crawling of the pages by a robot, the webmaster would have been better advised to insert the necessary commands in the HTML code of the pages, so that robots cannot crawl the pages and indexed the content. Just inserting such prohibition of crawling in terms and conditions which the user may access to but which he does not have to actually accept before continuing on his visit of the site does not guarantee the enforceability of the clauses.

  An exception would be the case of a website designed in such a way that the user could not pretend that he has not seen, neither read and accepted the terms and conditions. However this will probably only happen in a limited number of cases.

  On web data, **only intellectual property laws in general** (and in particularly, mainly copyright law) will govern what can be made (i.e. copyright for the use of the works, and database protection legislation for the use of the data, in case there actually is a database).

Levels 2, 3 and 4 - Social networks ("many to many"), Contractual Data ("one to many"), Confidential Data ("one to one")

[42] See however the recent study commissioned by the European Commission on the legal protection of trade secrets and confidential information.

- As opposed to what is happening with web data, it is our understanding that the level of access to social, contractual and confidential data[43] is twofold: access to such data are initially subject to **contractual clauses** agreed to by the parties (with such clauses being more or less restrictive, depending on the level of access – from not very restrictive (level 2) to very restrictive (level 4), and, **as a second layer, to intellectual property laws for the remainder**.

  It is thus only as regards these categories of data that the interplay between intellectual property rights and contractual clauses will have to be analysed and may have an impact, by either going further than what IP law would impose on users or on the contrary by relieving users from obligations which would otherwise derive from IP (be it copyright or database legislation). We will discuss this when dealing with the question whether a TDM exception could be overridden by contract or not (Part IX).

c) Some additional models: reuse of public sector information, Open Access, and Creative Commons

When describing the different modes of access to data which are relevant for data analysis, we think three additional models need to be presented: (i) access based on the existing EU legislation on reuse of public sector information ("**PSI**"), and (ii) the growing trend of publishing works in open access ("**OA**") or under Creative Commons ("**CC**") licenses.

We will present these three themes hereunder and explain each time where they fit in the four categories described above and why we consider that they are relevant for TDM. Each of them does facilitate access to data, which in turn should facilitate data analysis.

(i) Reuse of public sector information (PSI)

As part of the "Digital Agenda for Europe", the European Commission has set out a list of actions to be achieved in 2020; action 3 proposes to "open up public data resources for re-use".

In that context, the European Commission was referring to "public data" as:

> *"all the information that public bodies in the European Union produce, collect or pay for. This could include geographical data, statistics, meteorological data, data from publicly funded research projects, and digitised books from libraries. We speak about "open" public data when the data can be readily and easily consulted and re-used by anyone with access to a computer".*[44]

The re-use by the private sector of public data has been promoted so far in the EU through Directive 2003/98 and the recent Directive 2013/37 (often referred to as "the PSI Directives") and we think it is relevant to briefly mention them in this Study as this may have a positive impact on some sector-specific TDM activities, i.e. when the projects aim at analysing (mining) data held by public administrations.

The Directive 2003/98 on the re-use of public sector information ("PSI I Directive")[45] establishes a minimum set of rules governing the re-use and the practical means of facilitating reuse of existing

[43] Because they fall outside of the scope of the present Study, confidential data will not be studied in details.

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

[45] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information, *JO* L 345/90.

documents held by public sector bodies[46] of the Member States (Article 1). It ensures a level playing field for commercial re-users of public data[47], which includes companies involved in data analysis.

Under the PSI I Directive, the "re-use" of public data refers to (Article 2): *"the use of documents[48] held by public sector bodies, for commercial or non-commercial purposes other than the initial purpose within the public task for which the documents were produced".*[49]

The PSI I Directive was very recently complemented by a PSI II Directive [50]. The PSI II Directive has opened up even larger possibilities to re-use (including by data analysis) larger areas of information held by the public sector, by:

- imposing more obligations on public administrations to deliver the information in digital formats;
- strictly regulating the price which administrations may ask for the making available of the information;
- enlarging the sectors progressively falling under the obligation to allow re-use (libraries, museums, universities – all these institutions however within clear limits and possibilities of derogation).

This obligation to make more and more documents re-usable is of course very interesting for data analysis on public data, because it enlarges the corpus of works available for data analysis.

The right to re-use public data is not absolute. The PSI Directives, when read together, exclude such re-use for reasons such as national security, commercial confidentiality, protection of personal data and, most importantly for our topic, documents on which third parties hold intellectual property rights. Indeed, one must keep in mind that the PSI II Directives do not provide for a new exception to the intellectual property rights owned by a third party when said third party's works are held by a public sector institution.

The right to re-use public data can be granted for free, or subject to payment. However, there are strict rules regarding the price which administrations may ask (which should, except in certain circumstances, remain limited to covering costs involved).

The right to re-use public data can be free (i.e. no conditions attached) or subject to certain terms of use (i.e. to a "license"). The PSI Directives however prohibit conditions which unnecessarily restrict possibilities for re-use and e.g. which restrict competition (Article 8 of the PSI Directive[51]). The PSI II Directive confirms this prohibition[52].

---

[46] "'Public sector body' must be understood in a broad sense as meaning (Article 2.1): "the State, regional or local authorities, bodies governed by public law and associations formed by one or several such authorities or one or several such bodies governed by public law".

[47] Article 3: "Member States shall ensure that, where the re-use of documents held by public sector bodies is allowed, these documents shall be re-usable for commercial or non-commercial purposes in accordance with the conditions set out in Chapters III and IV. Where possible, documents shall be made available through electronic means".

[48] Article 2: "document" means: "(a) any content whatever its medium (written on paper or stored in electronic form or as a sound, visual or audiovisual recording); (b) any part of such content".

[49] "The purpose of the law [concerning the reuse of public sector information] [...] is essentially economic and thus differs from the regulations on access to public sector information. After all, the first fits into the functioning of the internal market and exploitation of public resources, while the latter is primarily a human rights (literal translation from French) ». C. DE TERWAGNE and J-Ph. MOINY, « A la croisée de la publicité de l'administration, de la réutilisation des informations du secteur public et de la protection des données : exemple de la directive INSPIRE », CDPK, Vanden Boele, 2010, p. 127.

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

[51] Article 8 on "Licenses":

1. *Public sector bodies may allow for re-use of documents without conditions or may impose conditions, where appropriate through a license, dealing with relevant issues. These conditions shall not unnecessarily restrict possibilities for re-use and shall not be used to restrict competition.*

Issues of formats under which the data are made available by public administrations could cause problems. For that purpose, the rules imposed on the public authorities have been strengthened, so as again to facilitate the reuse by third parties of the documents in standard formats. The PSI II Directive goes further by requiring public sector bodies to make documents available in "open and machine-readable format together with their metadata"[53]. Moreover, "both the format and the metadata should, in so far as possible, comply with formal open standards". This can be certainly helpful for companies involved in data analysis activities. The *rationale* is explained in Recitals 20 and 21 of the PSI II Directive[54].

Where does PSI fit in our four categories? PSI falls either in the "all to all" category (when no conditions are attached to the reuse of the information) or in the "one to many" category (when certain conditions are being imposed to reusers – as explained, the administrative entity authorising the reuse may attach certain licence conditions to such reuse, within limits).

What is the relevance of PSI for data analysis? The definition of re-use given by the PSI Directives is very broad as it includes all kinds of uses of documents, including use for data analysis. Moreover, the purpose can be commercial or non-commercial. This opens the door for commercial data analysis of public sector information e.g. for marketing purposes. The PSI Directives provide a wider spectrum of opportunities than the exception for research in the Infosoc Directive (limited to "uses for the sole purpose of scientific research" and "justified by the non-commercial purpose to be achieved" – see *infra*). It is however of a more limited impact as it is a sector-specific regulation (i.e. information held by the public sector) It also has a specific logic based on the fact that the public sector and its activities are financed by public money.

Examples of sectors where the activities have developed thanks to the reuse legislation are data mining activities in the geographic information (GIS) sector[55]. There are also developments in e.g. the health and social security sector (for purposes of e.g. control of costs of social security).

In its "Digital Agenda for Europe", the European Commission mentions as examples GPS, weather forecasts, financial and insurance services [56].

---

2. In Member States where licenses are used, Member States shall ensure that standard licenses for the re-use of public sector documents, which can be adapted to meet particular license applications, are available in digital format and can be processed electronically. Member States shall encourage all public sector bodies to use the standard licenses.

[52] In Article 8, paragraph 1 is replaced by the following: "1. *Public sector bodies may allow re-use without conditions or may impose conditions, where appropriate through a license. These conditions shall not unnecessarily restrict possibilities for re-use and shall not be used to restrict competition".*

[53] According to Article 2 of the PSI II Directive, "machine- readable format", ""open format" and "formal open standard" mean:

> *"machine- readable format means" a file format structured so that software applications can easily identify, recognize and extract specific data, including individual statements of fact, and their internal structure;*
>
> *"open format" means a file format that is platform-independent and made available to the public without any restriction that impedes the re-use of documents;*
>
> *"formal open standard" means a standard which has been laid down in written form, detailing specifications for the requirements on how to ensure software interoperability".*

[54] *"(20) To facilitate re-use, public sector bodies should, where possible and appropriate, make documents available through open and machine-readable formats and together with their metadata, at the best level of precision and granularity,* ***in a format that ensures interoperability****, e.g. by processing them in a way consistent with the principles governing the compatibility and usability requirements for spatial information under Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).*

(21) A document should be considered to be in a machine- readable format if it ***is in a file format that is structured in such a way that software applications can easily identify, recognize and extract specific data from it.*** *Data encoded in files that are structured in a machine- readable format are machine-readable data. Machine- readable formats can be open or proprietary; they can be formal standards or not. Documents encoded in a file format that limits automatic processing, because the data cannot, or cannot easily, be extracted from them, should not be considered to be in a machine-readable format. Member States should where appropriate encourage the use of open, machine-readable formats" (emphasis added).*

[55] See e.g. the activities of the national geographic institutes, together with Eurogeographics – www.eurogeographics.com

Lastly, the directives on the **re-use** of data held by the public sector [57] should not be (yet are sometimes) confused with certain texts organizing **access** by citizens to information held by the public sector. Issues of access to administrative documents remain largely a competence of the Member States, apart from a few initiatives taken by the European authorities[58] and are of little or hardly no relevance for TDM, because they usually require that the applicant has a personal interest in obtaining such access, because such access will be limited to the information relevant to said person and because access will not *per se* allow re-use (and thus a priori not data analysis either). We only mention them to remind that the confusion should not be made between legislation on re-use of PSI and legislation on access to administrative documents.

(ii) Open Access and Creative Commons

Open access (OA) and Creative Commons (CC) do not stem from legislative intervention. Quite to the contrary, they originate from private initiatives, which are now slowly endorsed by certain private and public organisations in Europe and elsewhere (e.g. the White House uses CC for its website). It seems useful to mention them here, as a factor which may have an impact on the development of TDM and which should be taken into account, if policy decisions are to be made in the area of TDM.

OA, while initially of a voluntary nature, sometimes "enters" in the administrative/regulatory sphere and becomes a compulsory model when public authorities decide that they will subsidise e.g. university research on the condition that the results of said research are made available in open access: this is then one of the (binding) conditions of the grant of subsidies.

- **Open Access**

It goes beyond the scope of this Study to analyze the overall impact which the Open Access movement will have on TDM but it seems undeniable that it will facilitate TDM. Indeed, facts show that OA is bringing changes to the publishing industry landscape. In August 2013, a Report on the "Proportion of Open Access Peer-Reviewed Papers at the European and World Levels (2004-2011)" indicates that the tipping point for OA (more than 50% of the papers available for free) has been reached in several countries, including Brazil, Switzerland, the Netherlands, the US, as well as in biomedical research, biology, and mathematics and statistics[59].

There are mainly three types of OA publishing: Gold OA, Green OA and Hybrid OA:

- under the "*Gold OA*" model, the articles are made immediately and freely available for reading and to re-use by the reader. The reader bears no costs. The costs are generally shifted away from the reader to the author, its institution or its funding body;
- under the "*Green OA*" model, the articles are archived by the author – or a representative - in an online repository before, after or alongside its publication. It is generally the author's final peer-reviewed version (the accepted manuscript before it is prepared for publication), not the published version. Articles are often posted with an embargo period (of one up to several months depending of the publishing house internal policy). No contribution is made to the costs of publication;

[56] Digital Agenda for Europe: A Europe 2020 Initiative , http://ec.europa.eu/digital-agenda/en/pillar-i-digital-single-market/action-3-open-public-data-resources-re-use

[57] To be complete, one could also mention a Commission Decision which specifically deals with the re-use of data held by the Commission itself, i.e. Commission Decision of 12 December 2011 on the reuse of Commission documents (2011/833/EU)

[58] See Directive 2003/4 on public access to environmental information and repealing Council Directive 90/313/EEC; also, more indirectly of use to the private sector, Directive 2007/2 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

[59] ARCHAMBAULT, E., AMYOT, D., DESCHAMPS, Ph., NICOL, A., REBOUT, L., and ROBERGE, Gu. , « Proportion of Open Access Peer-Reviewed Papers at the European and World Levels – 2004-2011 », Report produced for the European Commission DG Research & Innovation, August 2013, p. 1.

- finally, under the "*Hybrid OA*" model, subscription-based journals allow authors to make individual articles available on open access upon payment of an article publication fee.

Many OA initiatives have emerged in the years 2000. For the purpose of this Study, we chose to focus on two main ones: the Budapest Open Access Initiative (BOAI)[60] in 2001 and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities in 2003 (the "Berlin Declaration"). Thousands of individuals and organizations across the world have already signed the BOAI and the Berlin Declaration[61]. Both initiatives aim at making *research free and available to anyone* with a computer and an internet connection. The term "*research*" is here broader than published works and must be understood as including all the documentation and the research (including raw data and metadata) behind the works[62]. The terms "*free and available to anyone*" mean that authors and right holders agree to grant to the public the right to use the protected works to the fullest extent permitted by law. Open access documents are not only available to the research community, but to the public at large. The only constraint shall usually be to let to the authors the control over the integrity of their work and the right to be properly acknowledged and cited[63].

Publishers sometimes say that allowing data analysis can create concerns or risks for the security of their platforms or block the access to their platforms due to the overload of queries which end up on their platforms. In OA models, this problem is less likely to be present since, in many cases, works made available in OA are placed on dedicated repositories (separate from the publishing platforms of the publishers), which are then dimensioned to accept large number of queries and download requests.

Where does OA fit in our four categories? OA falls either in the "all to all" category (when no conditions are attached to the reuse of the information) or in the "one to many" category (when certain conditions are being imposed to reusers).

Why is OA relevant for data analysis? OA is *per se* pro-TDM because it grants to the public (including commercial companies) the right to use protected works to the fullest extent permitted by law, thus including for data analysis purposes (and not only for scientific research). TDM would only be prohibited if it is undertaken against national security or public order (and for any other acts not permitted by law). The rights of reuse granted to readers under OA schemes include the right to undertake data analysis on the documents published. So, all content made available under OA schemes is in principle free for data analysis. "Open access" is however not a protected trademark or label, so that it may still be that some schemes of distribution of content are presented as being "open access" yet do contain certain

[60] http://www.budapestopenaccessinitiative.org/

[61] To date, the BOAI has been signed by 5767 individuals and 678 organizations, and the Berlin Declaration has been signed by 469 organizations.

[62] The BOAI and the Berlin Declaration define *research* as: "peer-reviewed journal articles, but it also includes any unreviewed preprints that [scholars] might wish to put online for comment or to alert colleagues to important research findings." It does not include books from which their authors would prefer to generate revenue. It does not include any non-scholarly writings, such as novels or news" (BOAI) . And "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material" (Berlin Declaration).

[63] *"free availability to their works on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.* ***The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited****" (BOAI). And ""To grant(s) to all users a free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose,* ***subject to proper attribution of authorship*** *(community standards, will continue to provide the mechanism for enforcement of proper attribution and responsible use of the published work, as they do now), as well as the right to make small numbers of printed copies for their personal use. A complete version of the work and all supplemental materials, including a copy of the permission as stated above, in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards (such as the Open Archive definitions) that is supported and maintained by an academic institution, scholarly society, government agency, or other well-established organization that seeks to enable open access, unrestricted distribution, inter operability, and long-term archiving" (Berlin Declaration).*

restrictions which entail limitations for data analysis; we did not make a survey of the different OA schemes existing or progressively developing.

- **Creative Commons**

While OA concerns scientific articles and publications, Creative Commons (CC) licenses may apply to all types of works (scientific or not). As is well known, there are different CC licenses, some more permissive than others. Some types of licenses are particularly broad: the public domain dedication mark such as CC0 ("No Rights Reserved")[64] (yet its validity is doubtful under certain national copyright legislation) or PDM ("No Known Copyright"[65]) – also called CC Zero – and the license CC-BY ("Attribution")[66], the license CC-BY-SA ("Attribution/ShareAlike")[67] or the ODbL license ("Open Data Commons Open Database License"[68]).

By contrast with traditional publishing where users need to negotiate individual licenses for each use of the works, CC publishing does not discriminate between the beneficiaries of the license, the permitted uses (all uses permitted by law – except commercial uses in some instances), the price (CC means free access and use), and the duration of the licenses. Negotiating individual licenses is time-consuming and costs money, whereas CC publishing gives full access and use of the works at no costs.

While formats in which works are published may cause concerns for data analysis purposes, the existence of various CC or similar "open licenses" can also create "compatibility issues" (much like in the software open source licenses – and the same problem exists with OA publishing). While the purpose is to facilitate further dissemination of protected works, the result is sometimes that authors/re-users do not know exactly what they can or cannot do. One may only wish that more standardisation will progressively arise in this forest of licenses.

Where do CC fit in our four categories? CC fall in the "one to many" category (since conditions to the reuse are provided for in the various CC licences). However, depending on the CC licence, some will in fact be close, in practice, to the "all to all" category, e.g. with the "BY" CC licence (where the only condition is to mention the name of the author).

Why are CC relevant for data analysis? Publishing works under some of the most permissive CC licenses entails that the works may be reused, copied, adapted, etc. including for data analysis purposes. As the number of works made available on the Internet under such types of licenses grows, so do the possibilities to include them in TDM projects without having to consider whether copyright exceptions need to be invoked or not. Let us not forget however that some CC licenses will not allow commercial reuse of the works (CC NC licenses, with "NC" meaning "non-commercial"). Creative Commons have, for TDM, the advantage that they are usually used for non-scientific works. TDM projects which use as materials non-scientific works, such as data analysis for linguistic research, may therefore benefit from the growing use of CC licenses.

[64] "CC0 enables scientists, educators, artists and other creators and owners of copyright- or database-protected content to waive those interests in their works and thereby place them as completely as possible in the public domain, so that others may freely build upon, enhance and reuse the works for any purposes without restriction under copyright or database law". http://creativecommons.org/about/cc0

[65] "Our Public Domain Mark enables works that are no longer restricted by copyright to be marked as such in a standard and simple way, making them easily discoverable and available to others". http://creativecommons.org/about/pdm

[66] "This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation". http://creativecommons.org/licenses/

[67] "This license lets others remix, tweak, and build upon your work even for commercial purposes, as long as they credit you and license their new creations under the identical terms". http://creativecommons.org/licenses/

[68] http://opendatacommons.org/licenses/odbl/

III. ACTS AND EXCLUSIVE RIGHTS

In this Part III, as invited by the Terms of Reference, we will "*assess what acts and corresponding rights could be relevant for text and data mining activities, e.g., the reproduction right provided for in Article 2 of the Infosoc Directive and in Article 3 a) of the Database Directive, as well as the sui generis right provided for in Article 8 of the Database Directive*".

For the purpose of this chapter, we will first describe briefly the different steps of a normal data analysis process; we will then assess what acts and corresponding rights, as defined by the Infosoc Directive and the Database Directive, are (or not) relevant for data analysis activities.

A. HOW DOES DATA MINING WORK?

There are many ways to mine a set of data. There are also different ways to describe the process from a technical point of view; we tried to base our analysis on a general description. It may be that deviations are possible, and other descriptions exist. Furthermore, technology evolves rapidly and may entail changes to the description.

We will present here different steps that can be involved in a mining process. It is not the objective to give a detailed description of the process but only to recap whatever is necessary for the understanding of the Study. Our analysis below of the different steps of a data analysis is mainly inspired from the "ETL" technique, meaning: **E**xtract, **T**ransform, and **L**oad. This technique is commonly used to integrate data from multiple applications and sources. It is important to mention that, if ETL is probably one of the most common methods, there are other techniques (such as "ELT" – see hereunder) which do not always correspond to an ETL mechanism.

The choice of the mining technique depends on the goal and the conditions of the data analysis (technological context, quality of the hardware, costs, choice of software, time to achieve the data mining, etc.).

The different steps can be described as follows:

1. Individual content is **extracted** from outside sources (or sometimes **created**) – we assume here that it falls under copyright protection or database protection. We will call this phase "Obtaining the sources";
2. Content is, when necessary, **transformed** to fit operational needs;
3. Content is **loaded** into a data set, repository or collection;
4. Data miners gain access to the data and the mining (**analysis**) tools are applied to the data set;
5. New knowledge is created as a result of the analysis (usually a **report** can be drafted).

One could also consider that steps 1 to 4 concern the data mining **process** itself, while step 5 (drafting a report) could be considered more as the **result** (or "the **output**") of the mining process. We will come back on the distinction between the process and the result later in this Study. In most cases, while being the result of the data mining process, the report will not contain or display any of the data that have been "mined" and it is by no means a "summary" of all pre-existing outsources/data.

All those steps are not necessarily involved in a mining process. And as we already said, there are other methods to mine data, for example the "ELT" method (**E**xtract, **L**oad, **T**ransform), where the order of the steps is reversed. The difference between the two approaches (ETL and ELT) lies in the place where the processing happens: while the ETL processing of data happens in the ETL tool (transformation of the data in a cache or RAM memory, the data being loaded afterwards), the ELT processing of data happens

in the database engine.[69] The steps of the ELT are thus 1. extraction, 2. loading in a database and 3. transformation of the data. From a legal point of view, the analysis regarding the rights being involved (and whether or when e.g. a reproduction takes place may be different according to the technique being used.

B. WHAT ACTS AND CORRESPONDING RIGHTS ARE RELEVANT?

We shall hereafter examine which exclusive rights, both in the InfoSoc Directive and in the Database Directive, may come into play in the TDM steps described above.

a) The reproduction right in the InfoSoc Directive

One should distinguish copyright applying to (1) **works** being copied as part of the data analysis process, and copyright applying to (2) the **selection and arrangement** of such works or data (with such works or data being protected or not by copyright).

Technically speaking, it is often considered that data analysis involves, at some stage (particularly in steps 2 and 4 mentioned above), the copying of all or part of the data under investigation.

This view transcends the exceptions currently in force or under discussion in Japan, the UK and Ireland. In all these texts, it is taken for granted that data analysis does involve some form of copying.

> "*For the purpose of information analysis […] 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*". **Article 47 of the Japan Copyright Act (our emphasis)**
>
> "*Where a person has lawful access to a copy of a copyright work, copyright is not infringed where that person **makes a copy** of the work for the purposes of carrying out an electronic analysis of anything recorded in the work*". **Draft Section 29A of the UK Copyright, Designs and Patents Act[70] (our emphasis)**
>
> "*It is not an infringement of the rights conferred by this Act for a person **to reproduce** a work for a purpose to which this section applies if […] (2) This section applies to — (a) text-mining, data-mining, and similar analysis or research*". **Draft Section 106F of the Irish Copyright and Related Act** (**our emphasis**)

Authors and researchers generally (but not always) agree on this issue:

> "*Due to the fact a computer must make a copy of an entire copyright work in order to perform the same activity, the process of data mining becomes subject to copyright law*". **Universities UK and UK Higher Education International Unit[71]**

[69] Stackexchange.com, "Questions", "What are the arguments in favor of using ELT process over ETL?", http://dba.stackexchange.com/questions/19242/what-are-the-arguments-in-favor-of-using-elt-process-over-etl

[70] On the same subject: "Copyright law was not established to regulate the deployment of analytic technologies, but because these technologies involve acts of copying, copyright does take effect in this area, even where the products of the technologies do not contain protectable expression or affect the markets for the primary works". UK IPO Impact Assessment, "Exception for copying of works for use by text and data analytic", December 2012, http://www.ipo.gov.uk/consult-ia-bis0312.pdf

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

> "*Automated text processing presents a paradox for copyright law*[72]*. On one side, automated processing presupposes the repeated copying of whole works; in this respect, it is an exemplary prima facie case for infringement. On the other side, however, the purpose of this reproduction is to extract information from texts and about texts, an activity that does not normally amount to an infringement in copyright law*". **Maurizio Borghi and Stavroula Karapap**[73]

Some authors deviate from this general trend:

> "*In the interest of a general analysis, it will be assumed that there is actual copying of substantial of contents during the mining operation, although it is understood that this may not always be the case*". **Andres GUADAMUZ and Diane CABELL**[74]

Moreover, with regard to data analysis applied to text, text documents typically provided by publisher and/or repositories are in PDF format. However, PDF is a major barrier to data analysis and must, therefore, be converted (i.e. copied) to XML format to work properly[75]:

> "PDF is evil". **Cliff Lynch**[76]

> "*Once the documents are collected, it is not uncommon to find them in a variety of different formats, depending on how the documents were generated. […] Clearly, if we are to process all the documents, it's helpful to convert them to a standard format. The computer industry as a whole, including most of the text-processing community, has adopted XML (Extensible Markup Language) as its standard exchange format*". **Sholom M. Weiss, Nitin Indurkhya and Tong Zhang**[77]

The act of copying is considered as an exclusive right by Article 2 ("Reproduction right") of the Infosoc Directive:

> "*The Member States shall provide for the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part: (a) for authors of their works"*.

The right of reproduction provided hereto must be interpreted broadly and "extends to every act of reproduction, however transient or irrelevant it may be from an economic or functional perspective (subject to the exception for temporary reproduction provided for in Article 5.1 of the Infosoc Directive – we will analyse this in Part IV of this Study)"[78].

Recital 21 of the Infosoc Directive provides that:

---

[72] "Automated text processing first gained copyright relevance in the context of Google Books. […] The [Google Books Settlement] allowed "qualified users" to request access to the corpus of Google Books in order to conduct uses for "non-consumptive research", defined as "research in which computational analysis is performed on one or more Books, but not research in which researcher reads or displays substantial portions of a Book to understand the intellectual content presented within the Book. This included textual, linguistic, and image analysis, text and information extraction, automated translation, and indexing and search. All these activities involve wholesale systematic and repeated reproduction of large numbers of works. The bigger the volume that is fed into computers, the better the results that can be produced. It is in this respect that automated processing become qualitatively and quantitatively relevant from a copyright perspective". BORGHI, M., and KARAPAPA, S, op. cit., note 6, p. 51

[73] BORGHI, M., and KARAPAPA, S, op. cit., note 6, p. 51

[74] GUADAMUZ, A., and CABELL, D., op. cit. note 32, p. 6.

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

[76] Clifford Lynch has led the Coalition for Networked Information (CNI) since 1997. More information: http://www.cni.org/about-cni/staff/clifford-a-lynch/

[77] WEISS, S. M., INDURKHYA, N., and ZHANG, T., *Fundamentals of Predictive Text Mining*, Texts in Computer Sciences, Springer, 2010, p. 15.

[78] WALTER, M. W., and VON LEWINSKY, S. V., *European Copyright Law :A Commentary*, Oxford University Press, 2010, pp. 967and 968.

> "*This Directive should define the scope of the acts covered by the reproduction right with regard to the different beneficiaries. This should be done in conformity with the acquis communautaire. A broad definition of these acts is needed to ensure legal certainty within the internal market*".

In the *Infopaq* case, the European Court of Justice ruled that:

> "*An act occurring during a data capture process, which consists of storing an extract of a protected work comprising 11 words and printing out that extract, is such as to come within the concept of reproduction in part within the meaning of Article 2 of Directive 2001/29 on the harmonisation of certain aspects of copyright and related rights in the information society, if the elements thus reproduced are the expression of the intellectual creation of their author; it is for the national court to make this determination*".[79]

Michel M. WALTER and Silke VON LEWINSKY further explain:

> *"The definition as laid down in Article 2 of the Information Society Directive combines all clarifying elements as set out in the existing international treaties and European Directive so far adopted. Its wording as well as the concern laid down in Recitals 9, 10, and 21 implies that the right of reproduction is to be construed broadly. The European Court of Justice in its Infopaq Judgment explicitly confirmed this view and held that the protection conferred by Article 2 must be given a broad interpretation. […] The notion of reproduction is to be understood in a broad sense and is determined technically rather than functionally, in order effectively to protect the author against the use of his works by third parties without his consent. Hence, the reproduction right according to Article 2 of the Information Society Directive extends to every act of reproduction, however transient or irrelevant it may be from an economic or functional perspective. This concept results in the mere use of a work being considered subject to the author's consent, whenever such use necessitates reproduction.*[80]

On that basis, we consider that data analysis **generally involves an act of copying** – in accordance with Article 2 of the Infosoc Directive - of whole or part of the data being processed during the data analysis activities.

On condition however that the works being processed are protected by copyright, which may not in all cases be the situation.

This conclusion might not apply in certain exceptional circumstances, i.e. if it happens that, through the process of data analysis, the software only "crawls" through the text or the data, and processes them "one by one", without copying the whole text but only one data or word or just a few of them at a time. In that case, even if one does not ignore that the ECJ in its Infopaq case accepted that an excerpt of a sentence made of 11 words could be protected by copyright, it would be our understanding that, in such a case where the software only "swallows" one or two words or pixels or data or sounds at a time and then goes on to the next ones without keeping a copy of them but just "counting" the number of occurrences of, say, the word "malaria", then we would argue that **no copying relevant in terms of copyright takes place** and that such activity does not require the consent of the rightholder – and thus no exception is needed.

b) The other exclusive rights in the InfoSoc Directive or in copyright

Apart from the reproduction right, are there other rights which are relevant in the InfoSoc Directive or generally under copyright and for which it may be necessary that a license be obtained from the rightholder before the data analysis may take place?

This is not our view. TDM will require processing the works in different ways (like switching formats, separating certain data from the rest of the texts, isolating some parts, etc.). Even if the adaptation right

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

[80] WALTER, M. W., and VON LEWINSKY, S. V., op. cit., note 58, pp. 967 and 968.

and the translation right are not explicitly mentioned in the InfoSoc Directive[81], the concept of the reproduction right in the InfoSoc Directive has intentionally been made very broad ("*the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part*"), so that the technical operations and processing to be performed on the works in the context of a data analysis process would in our view fall under the broad concept of "reproduction".

The case of transformation in another format (like from PDF to HTML or HML) will easily qualify as a reproduction.

It is maybe less certain for a translation from one language to another (from French to English, for instance). There, maybe a difference should be made between a *human* translation (where choices are made by the translator, who will own a copyright on his translation because he will have made a new original work)[82], and an *automated* translation by a software (where no human intervention is needed at the level of the translations and where one could convincingly argue that the translation by the machine is a mere reproduction in the sense of the InfoSoc Directive). In both cases (change of format and translation), if the files are still protected by copyright, such transformation into another technical format will require the authorization of the rightholder (unless an exception can be invoked – which will be analyzed in Part IV of our Study).

But let us imagine that format switching and translation in another language (other than by machine) are considered by some as adaptations more than as reproductions. Does that imply that, should the Commission want to consider a TDM exception, some could argue that the Commission has no jurisdiction to do this on the basis of the *acquis* because the adaptation or the translation rights have not been harmonized in the InfoSoc Directive? We do not think so, for the following reasons:

- First, adaptations (and all arrangements and modifications) have been harmonized in the Database Directive; it was also done in the Software Directive – so, the idea of a harmonization of the adaptation right is not an heresy as such and should not be considered as a "*chasse gardée*" which *per se* and by its very nature would escape the jurisdiction of the Commission;
- Secondly, and more importantly, in the Infosoc Directive, some exceptions have been introduced which are exceptions not only to the reproduction right but also to the adaptation right, even though the adaptation right has not been harmonized – and this has not met any significant opposition:
    - this is clearly the case for the parody (caricature and pastiche) (which should reasonably considered more as an adaptation than as a reproduction);
    - some other exceptions will in many cases involve automatic changes of format – and nobody argued that this should qualify as an adaptation and therefore could not be harmonized under the umbrella of the reproduction right; a clear example of this is the reproduction for people with a disability (requiring translations in Braille, adaptations of format, etc.): did anyone oppose this because adaptation and translation were not harmonized by the Directive? (no);
    - use of an artistic work in the form of a building or a drawing or plan of a building for the purposes of reconstructing the building (article 5.2.m): going from 2 dimensions to 3 dimensions can arguably be considered not just as a reproduction but also as an adaptation;
- Third, one can probably argue that in any event all copies made in the context of TDM that may be considered as "adaptations" under national law are in fact acts of reproduction under the EU copyright rules and the InfoSoc Directive. Accordingly, there does not seem to be any risk of introducing an exception to rights that are not harmonised at EU level.

[81] Except if one would consider that the reproduction right also includes the adaptation right and the translation right (and in some countries, it does).

[82] As in the Italian expression: "***Traduttore, traditore"*** ("To translate is to betray" or literally, "translator, traitor"). This expression means that the translation of a text from one language into another can never fully comply with the text of the original work.

.

The other rights, i.e. the distribution right, the communication to the public right and the making available right are not relevant for our Study.

Indeed, one must make a distinction between (a) the **process** of data analysis itself and (b) acts concerning the production of an **output** from the data analysis:

(a) We do not see where a distribution (according to the InfoSoc Directive) would take place in the normal process of data analysis; amongst other things, the data analysis process will normally not involve that tangible copies be distributed to "a public"; the only recipients will be the persons (researchers or others) who will be involved in the analysis;

We do not see either where a communication "to the public" or a making available "to the public" would take place in the normal process of data analysis; again, such process does not involve that the texts, images, sounds to be analyzed are made available to "a public";

(b) As such, the data analysis process does not involve that the output resulting from the analysis be distributed to a public, communicated or otherwise made available to a public. It may indeed in many cases remain confidential or restricted to a limited number of recipients. Furthermore, and more importantly, as indicated above, the output itself, i.e. the result of the data analysis, which will typically be in the form of a report, of a graphic or of a table, does not amount to a reproduction, nor to a distribution, communication to the public or making available of the underlying elements which have previously been copied: in such output indeed, only results like statistics, number of occurrences, new relations, etc. will be presented. It may be that in such output a few words coming from the analysed texts (if the analysis concerns texts) are found in the output, but as we mentioned above regarding the reproduction right, this does not suffice to conclude that a distribution, a communication to a public or a making available of pre-existing copyright protected elements has taken place..

It is therefore probably not a coincidence that the draft legislation we referred to above do not refer to these other rights (except to a limited extent to the adaptation right): the reason is that there is no need whatsoever to provide for an exception where the right is not exercised during the data analysis process nor to produce the analysis output.

c) Copyright in the Database Directive

(i) Why is this relevant?

For there to be a database, one must have the possibility to take some elements away without the whole (the database) losing all its use and significance (on the contrary, taking words out of a book will make the book incomplete and of little value to the reader). So, the Database Directive is only relevant when the activity of data mining concerns databases, not if it only concerns books or pieces of music taken in isolation, be it even in an electronic format. Indeed, not every file is a database, not every article is a database, not every book is a database. As recalled by the Database Directive (article 1.2), for the purposes of this Directive, 'database` shall mean a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means". It has largely been accepted that a fictional book as such, a painting, a picture, a film, is normally not a database.

In many cases, data mining will need to analyse large amounts of texts, sounds, facts, data, etc.

It may be that the whole corpus being analysed will itself not qualify as a database, because the elements might not be "arranged in a systematic or methodical way" nor "individually accessible by electronic or

other means": as was just said, a large book, for instance, is not a database; even a series of books (a medical encyclopaedia) might not always be considered as a database.

But it is highly likely that in many projects the corpus which will be subject to data analysis will itself comprise a number of databases (e.g. collections of scientific publications).

Many databases are not protected by copyright, due to a lack of originality. However, some databases are, so that we first need to examine whether data analysis raises a problem if these copyright protected databases are being analysed. Secondly, as shall be explained, even when the database itself is not protected by copyright, it may be that its content is protected by the *sui generis* right.

(ii) Protection of the database by copyright

As regards a possible protection of the database by copyright, the Database Directive provides as follows:

> *"In accordance with this Directive, 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".*

If the database is protected by copyright, the act of copying will apply to the selection or arrangement of elements contained in databases.

If the database is protected by copyright, the following consequences will apply, in accordance with article 5 of the Database Directive (Restricted acts):

> *In respect of the expression of the database which is protectable by copyright, the author of a database shall have the exclusive right to carry out or to authorize:*
>
> *(a) temporary or permanent reproduction by any means and in any form, in whole or in part;*
>
> *(b) translation, adaptation, arrangement and any other alteration;*
>
> *(c) any form of distribution to the public of the database or of copies thereof (...);*
>
> *(d) any communication, display or performance to the public;*
>
> *(e) any reproduction, distribution, communication, display or performance to the public of the results of the acts referred to in (b).*

(iii) Reproduction of a database

Article 5 a) of the Database Directive ("Restricted acts") is applicable to data mining and prohibits the copying of whole or part of the protected databases:

> *"In respect of the expression of the database which is protectable by copyright, the author of a database shall have the exclusive right to carry out or to authorize: (a) temporary or permanent reproduction by any means and in any form, in whole or in part"*

One has to make a distinction between the data mining process and the data mining output(s):

- during the data mining **process**, it may be that the selection or the arrangement of the data base are copied; this will for instance be the case if all data included in the database are copied: in that case, the selection is being copied; in other cases, parts of the arrangement could also happen to be copied during the process; this will however not necessarily be the case and will depend on the circumstances and the techniques being used;

- concerning the data mining **output(s**), we consider it most unlikely that the data analysis output will contain whole or part of a protected database which can be recognizable. This is because the data

analysis output generally is an independent creation and the differences with the original database are so important that the global impression is not the same.

B. MICHAUX explains that:

> "*Like every other authors, the author of a database shall have the exclusive right of reproduction. The reproduction, to fall within the scope of the right of reproduction of the author, does not have to be total. It can be partial as long as it borrows from the originality of the work. For that, the part of the database must be recognizable as such in the alleged infringement. It can happen that the derived work shows differences with the original work so that the similarities are not identifiable anymore. More precisely, from the moment the differences are so that the global impression is not the same, there is no infringement, but independent creation".*[83]

One could indeed follow the reasoning and consider that the output of the data analysis is so different, from a selection and arrangement point of view, that the elements on which copyright could have a grasp have become undetectable.

From hundreds or thousands of data or works which will have been analysed, indexed, compared, linked, aggregated, merged, disassembled, etc., it may be very hard to prove that the data come from a particular database and, more importantly, that they infringe upon the selection or arrangement of the database (which is what copyright protects in a database). Many databases will, in many cases, have been merged together, data may come from one or from another database (so that the selection element will hardly be of any relevance), the existence of one particular data in the final output will generally be impossible to be traced back and attributed to one particular database which has been mined. So the possibility to claim and prove infringement is highly unlikely. Also, the structure of the initial databases will hardly be found back in the final output, where so many operations will have taken place that the structure and arrangement itself of one particular database will be impossible to find (subject maybe to using special identification techniques like watermarking).

(iv) Translation, adaptation, arrangement and any other alteration of a database

One should again distinguish between the process and the output:

- The **process** of data analysis may involve an adaptation or arrangement of a database, but not necessarily; this will depend on the circumstances;
- The **output** will in our view in most cases be too distant from the databases which have been analysed and thus will not amount to an adaptation or alteration thereof.

With regard to the act of adaptation, translation or any arrangement of the database (article 5.b of the Database Directive), it is clear that, because data analysis may cover hundreds or thousands of copyrighted works, isolating the original ownership of those works may be impossible in practice. Moreover, we think that the acts of copying of the data very often happen at the stage of processing of the data (see our comment *supra*), and that the copied data are not identifiable anymore in the data analysis output:

> "*[T]here is an interesting question about being able to isolate any one publisher's work included in any particular text mining output. Though the publisher's server may have been interrogated by the text mining software, connecting the results of the mining process back to any one of the original information sources may prove difficult. Multiple results may have been derived from a wide variety of text sources and how can one be give credit for any one item? Computers are*

[83] « *La reproduction, pour relever du droit exclusif de l'auteur, ne doit pas être totale. Elle peut n'être que partielle pourvu qu'il y ait un emprunt à ce qui fait, en tout ou en partie, l'originalité de l'œuvre. Encore faut-il que la partie reprise soit reconnaissable comme telle dans la contrefaçon prétendue. Il se peut en effet que l'œuvre seconde présente des différences avec l'œuvre première au point que la similitude ne soit plus identifiable. Plus exactement, lorsque les différences sont telles que l'impression globale n'est plus la même, il n'y a pas contrefaçon, mais création indépendante* ». MICHAUX, M., *Droit des bases de données*, Kluwer, 2005, p.119.

> *logical, not creative. Computation using text mining to create a derivate work is essentially a mechanical activity. Derivative works can therefore be based on hundreds or thousands of separate copyrighted works. Isolating the original ownership of an idea or wording may be impossible. Text and data mining also needs to encompass the creation of extracts, translations and summaries of developments in various fields. Some of these derivative works are mechanically produced but others, such as creating a translation, still need elements of human creativity. So much so that copyright may be vested in the derived translation. There is an outstanding legal question of who can determine what is included from whom (copyright owner) in a newly derived work*?" **ICSTI**[84]

One issue which may deserve further consideration is the requirement under the Irish draft legislation to acknowledge the source where this is not impracticable (which we would guess it will be). This issue will be discussed when analysing the conditions and limitations which a legislator might impose before allowing data analysis without the need for a license (Part IX).

Regarding adaptation, article 47 of the Japan Copyright Act provides:

> "*For the purpose of information analysis […] 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*".

From this article, one may conclude that in the legislator's opinion data analysis may involve, at some stage, the adaptation of the database being processed. One example could be the translation of the whole content of the database for the purpose of processing. This assumption has not been verified anywhere in the literature we have consulted so far. Adaptation may also sometimes be understood as shifting from one format to another; this might be what is here at stake and what the Japanese legislator had in mind.

As a preliminary conclusion, it seems that:

- Regarding the **process**: some data analysis operations will indeed amount to an adaptation, translation or arrangement of a database: this will particularly be the case where all the data (and thus, the selection made by the database maker) are being translated or transformed into another format): the selection itself is being copied also and may need to be somewhat adapted later on for the purposes of format shifting. The same goes for the structure: it may need to be entirely copied and then later one adapted or in some way rearranged to comply with the technical requirements of the analysis software. In these cases, not only has there been a reproduction of the database itself but also an adaptation thereof;
- Regarding the **output** however of a data analysis, it will in all likelihood not be considered as an adaptation of the database which has been analysed.

(v) Communication to the public of the database

Let us distinguish again between the process and the output:

- Concerning the data analysis **process**, there is no communication "to the public" if a database is being analysed (mined) by a group of researchers or a private company or any other "miner". A communication "to the public" does require that the work (here, the database) be communicated to a sufficiently significant number of persons;
- Concerning the data analysis **output**, it is very unlikely, in our view, that it will involve a communication to the public of the copyright protected elements of the database itself (i.e. the original selection or arrangement).

[84] International Council for Scientific Information (ICSTI), "Text and Data Mining", July 2009, p.12.

Considering that the output (which will in many case be a new text, a graphic, a chart, a design, a matrix, etc.) is an independent creation, the original data (underlying maybe the new work) will themselves arguably not be communicated to the public via this channel, in the same way as the knowledge gained by a researcher while preparing his thesis will in itself not be found as such in the thesis (at least not in a manner which infringes intellectual property rights of other rightholders).

It is indeed our view that when the output of the analysis is being communicated, such output will be composed of statistics, new relationships, new patterns, which by definition were not visible in the pre-existing works. Only really tiny elements might, by mistake or coincidence, escape the attention of the drafters of the output and appear as such in the presentation of the output.

The right of communication to the public includes, as we know, the making available to the public of their works in such a way that members of the public may access them from a place and at a time individually chosen by them. Here again, the data mining output will normally not give the possibility to access the protectable elements (i.e. the selection or the arrangement of the database). Only possibly some distinct elements of the selection or of the arrangement of the initial database may by chance appear in the output and be made visible and distinct from the rest to the observers, but this would definitely be an exception.

The fact that no communication to the public is involved in a normal data analysis process nor in outputs probably explains why the Japanese legislation and the draft UK legislation do not mention such rights and do not provide for an exception to them for data mining purposes: such exception is not necessary.

d) The sui generis right in the Database Directive

(i) Protection of the database by the sui generis right

The maker of a database will be granted *sui generis* rights (article 7.1. of the Database Directive) if he can show that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents of the database. Such *sui generis* rights will then allow the maker 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

First of all, it is important to point out that the database right (the *sui generis* right) exists regardless of the existence of copyright protection in the database or in its content, as the exclusive rights given to the database maker are separate from those arising from copyright. Article 7.4 of the Database Directive expressly provides that:

> "*The right provided for in paragraph 1 shall apply irrespective of the eligibility of that database for protection by copyright or by other rights. Moreover, it shall apply irrespective of eligibility of the contents of that database for protection by copyright or by other rights. Protection of databases under the right provided for in paragraph 1 shall be without prejudice to rights existing in respect of their contents".*

The *sui generis* right is made of two components, the extraction right and the reutilisation right.

According to article 7 of the Database Directive:

> *"Member States shall provide for a right for the maker of a database which shows that there has been qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents to prevent* ***extraction*** *and/or* ***re-utilization*** *of* ***the whole*** *or of* ***a substantial part, evaluated qualitatively and/or quantitatively****, of the contents of that database."*

We shall examine these two exclusive rights separately.

(ii) The extraction right

With regard to the act of extraction, it is unanimously admitted that the act of extraction must be understood broadly; it is prohibited to the extent that it applies to all or to a substantial part of the contents of the database.

What does "extraction" mean?

Article 7.2 a) of the Database Directive defines the act of 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*";

The Court of Justice of the European Union gives a broad definition of "extraction":

> "§ 51. *The use of expressions such as 'by any means or in any form' and 'any form of making available to the public' indicates that the Community legislature intended to give the concepts of extraction and re-utilisation a wide definition. In the light of the objective pursued by the directive, those terms must therefore be interpreted as referring to any act of appropriating and making available to the public, without the consent of the maker of the database, the results of his investment, thus depriving him of revenue which should have enabled him to redeem the cost of the investment*".[85]

The Court further notes that when a third party integrates data from a database into his own electronic system, he/she makes an extraction:

> "*Although William Hill is a lawful user of the database made accessible to the public, at least as regards the part of that database representing information about races, it appears from the order for reference that it carries out acts of extraction and reutilisation within the meaning of Article 7(2) of the directive. First, it extracts data originating in the BHB database by transferring them from one medium to another. It integrates those data into its own electronic system*. [86]

As we shall see in Part IV of this Study, there are *exceptions* to this exclusive right of extraction and there are also *rights* granted to legitimate users of a database.

We think that, in most cases, **the data analysis process will entail the extraction of all or a substantial part of the data held in the database** in order for such data to be processed for the purpose of "data analysis"; in this case, the authorization of the maker must be obtained. This is likely to be the case in many situations where databases are part of a corpus to be analyzed. It is no coincidence that many of the definitions we found of "text and data mining" do refer to "extracting data" or "extracting" information.

This being said, one may not pretend that data analysis *per se* involves the exclusive right of extraction of the maker: it could happen that the software only copies/extracts some relevant elements only of the database and that such elements are so small in number that they should be considered as unsubstantial.

It could also be, depending on the technology being used for the analysis, that no permanent or temporary "transfer" of data (whether of a substantial part or of unsubstantial parts) takes place: for instance, the software crawls through the data, does not copy the file nor even any data but "counts" occurrences or "registers a link between this data and another data": in such case, it seems that no

[85] CJE, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd*., par. 51.

[86] Court of Justice of the European Union, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd,* par. 65.

transfer takes place, and thus no extraction either. The maker of the database has then no ground to invoke his *sui generis* right.

(iii) The re-utilization right

With regard to the act of re-utilization, it is unanimously admitted that the act of re-utilization must also be understood broadly; it is prohibited to the extent that it applies to all or a substantial part of the content of the database.

Article 7.2 b) indeed defines the act of re-utilization as:

> "*any form of making available to the public of all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission*".

The Court of Justice of the European Union gave a broad interpretation of the concept of "re-utilization":

> "§ 51. *The use of expressions such as 'by any means or in any form' and 'any form of making available to the public' indicates that the Community legislature intended to give the concepts of extraction and re-utilisation a wide definition. In the light of the objective pursued by the directive, those terms must therefore be interpreted as referring to any act of appropriating and making available to the public, without the consent of the maker of the database, the results of his investment, thus depriving him of revenue which should have enabled him to redeem the cost of the investment*".[87]

The same Court decided that:

> "20. *that concept must, in the general context of Article 7, be understood broadly, as extending to any act, not authorised by the maker of the database protected by the sui generis right, of distribution to the public of the whole or a part of the contents of the database".*
>
> *"21. [Re-utilization] covers an act, such as those at issue in the main proceedings, in which a person sends, by means of his web server, to another person's computer, at that person's request, data previously extracted from the content of a database protected by the sui generis right. By such a sending, that data is made available to a member of the public*".[88]

Alain STROWEL and Jean-Paul TRIAILLE further explain that:

> "*If the right of extraction can be associated with the right of reproduction, the right of re-utilization could, in the same way, be associated with the right of communication to the public. The online transmission of part of the content of a database, be it protected or not by copyright, is subject to the authorization of the maker."* (literal translation from French).[89]

Benoit MICHAUX specifies:

> "[re-*utilization] refers to the diffusion of content, be it a distribution of physical materials incorporating the content of the database or a substantial part of it, or an online transmission or "any other form". Hence, it includes the communication to the public, by wire or wireless means,*

[87] Court of Justice of the European Union, 9 November 2004, Case C‑203/02, The British Horseracing Board Ltd and Others v William Hill Organization Ltd., paragraph 51.

[88] Court of Justice of the European Union, 18 November 2012, C-173/11, *Football Dataco Ltd and Others v Sportradar GmbH et Sportradar AG,* paragraphs 20 and 21.

[89] « Si le droit d'extraction peut être rapproché du droit de reproduction, le droit de réutilisation serait alors à rapprocher du droit de communication au public. La transmission en ligne d'une partie du contenu d'une base de données, qu'il soit protégé par le droit d'auteur ou non, est soumise à l'autorisation du fabriquant». TRIAILLE, J-P and STROWEL, A., op. cit., note 4, p. 278.

including the digital transmission. It includes the sale or any other transfer of property of tangible materials as well as their rental, but not the public lending" (literal translation from French).[90]

Let us again distinguish between the process of data analysis and the output(s):

Regarding the **process**: the same reasoning may be applied than for the right of communication to the public regarding the copyright protection of the database: there is no making available "to the public" (and thus no re-utilization) if a database is being analysed (mined) by a group of researchers or a private company or any other "miner". A making available "to the public" does require that the contents of the database be communicated to a sufficiently significant number of persons.

Regarding the **output**: as previously explained with regard to the right of communication to the public (above), there is a common understanding that the data analysis output is an independent creation, so that the contents of the database will as such arguably not be re-utilized / communicated to the public via this channel in any recognizable manner.

C. INTERMEDIARY CONCLUSIONS

We suggest referring to "data analysis" rather than to "text mining", "data mining" or "text and data mining", for reasons explained in our Study.

We explain why we consider that it does make a difference, from a legal and contractual point of view, whether the data are accessible under contractual terms or not; there are different levels of control organised by these contractual terms; the way by which they add restrictions which are not to be found in copyright law or by which they alleviate restrictions otherwise imposed by copyright law is important for the further development of data analysis as an economic sector.

We then try to demonstrate that:

- from a copyright standpoint:
  - when the data analysis is made on the basis of data and information which are protected by copyright, data analysis does, in most cases, involve a reproduction of protected materials and arguably also translating or adapting the same (with such translation or adaptation falling, in our view, within the scope of the reproduction right under the InfoSoc Directive), but not communicating them to the public nor making them available;
  - when the data analysis is made on the basis of data and information held in a database, it will only in some cases but not often involve a reproduction or an adaptation of the database itself and it will almost never involve a communication (or making available) of the database to the public;

- from a database protection law (*sui generis* rights) standpoint, when the data analysis is made on the basis of data and information held in a database :
  - data analysis will, in most cases, involve extraction of all or substantial parts of the contents of the database;
  - but it will normally not amount to re-utilizing the same.

[90] « [*La réutilisation] désigne une diffusion de contenu, que ce soit sous la forme de distribution d'objets physique incorporant le contenu de la base de données ou une partie substantielle de celui-ci, ou sous la forme d'une transmission en ligne ou sous « toute autre forme ». Elle inclut donc la communication au public, par fil ou sans fil, y compris la transmission numérique. Elle comprend la vente ou toute autre forme de transfert de propriété d'objets physiques ainsi que leur location, mais non le prêt public* ». MICHAUX, B., op. cit., note 52, p. 153.

IV. EXCEPTIONS AND LIMITATIONS

We will assess whether data analysis activities could be covered by the current exceptions to copyright and/or to the *sui generis* right.

We will cover, for copyright, the exception for temporary acts of reproduction in Article 5.1 of the Infosoc Directive, and the exceptions for scientific research in Article 5.3 a) of the Infosoc Directive and in Article 6.2 b) of the Database Directive; and for the *sui generis* right, the exception for scientific research in Article 9 b) of the Database Directive.

Despite the fact that the Terms of Reference do not explicitly refer to other exceptions, it is our understanding that Articles 6.1 and 8.1 of the Database Directive could be relevant for this Study. Indeed, Article 6.1 provides for an exception to copyright for the normal use of the "structure of the database" by the lawful user, while Article 8.1 refers to the rights of the lawful user of a database protected by a *sui generis* right to extract the data contained in the database.

Exceptions to copyright will be dealt under (A) hereunder; exceptions to the *sui generis* right under (B) hereunder.

A. EXCEPTIONS TO COPYRIGHT

We will deal with the exception for temporary acts of reproduction under (a), the exception for scientific research under (b) and (in addition to what was suggested by the Terms of Reference) the exception regarding the normal use of the database under (c).

a) Exception for temporary acts of reproduction

(i) The principles (reminder)

The InfoSoc Directive provides a list of exceptions and limitations to the reproduction right and to the right of communication to the public aiming to ensure a functioning internal market (art. 5 InfoSoc Dir).

Article 5.1. of the InfoSoc Directive. The exception allowing acts of temporary reproduction provided by art. 5.1 of the InfoSoc Directive reads as follows:

> *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.*

This provision has been examined extensively in the "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*"[91] and in the "*Study*

[91] DEPREEUW, S. and HUBIN, J.-B., "*Study on the territoriality of the making available right. Localisation of the act of making available to the public and its consequences*" *in* TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", European Union, October 2013, p. 109 et s., available on http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf. We repeated here the most important elements of their presentation but omitted the additional explanations on national case-law. For further information, please consult the full Study mentioned above.

to review options for developing the relationship between the reproduction right and the making available right in the context of the cross-border transmission of digital content"[92] by S. Depreeuw and J.-B. Hubin.[93]

We reproduce hereafter, with the authors' consent, a summary of the part of this study on the temporary reproduction.

The Court of Justice examined the conditions provided by art. 5(1) in *Infopaq I* and *II*, as well as in *Premier League*[94].

The Court first declared that the exception has to be interpreted strictly, being a restriction to the exclusive right of the author[95] and cumulative conditions have to be met in the sense that non-compliance with any one of them will lead to the act of reproduction not being exempted[96]:

1. the temporary copy must be transient or incidental;
2. it has to be an integral and essential part of a technological process;
3. its sole purpose must be:
    a. to enable a transmission in a network between third parties by an intermediary;
    b. or a lawful use of a work or protected subject-matter;
4. the act must have no independent economic significance.

We will successively analyse these conditions and how these have been interpreted by the ECJ. In our general Study on the Directive, we also analysed how national courts interpreted these conditions; we shall not reproduce this part in this Study[97].

According to the ECJ, the exception allowing acts of temporary reproduction provided by art. 5.1 has specifically been drafted to allow and ensure the development of new technologies and safeguard a fair balance between the rights and interests of right holders, on the one hand, and of users of protected works who wish to avail themselves of those new technologies, on the other[98].

Based on a technology-neutral approach similar to the one used to define the reproduction right, this exception mitigates the consequences of the broad definition provided by art. 2 of the InfoSoc Directive that includes "*temporary reproduction by any means and in any form*". Indeed, it was found necessary to allow certain copies forming a part of a technological process, such as copies needed to enable browsing or caching (rec. 33).

---

[92] DEPREEUW, S. and HUBIN, J.-B., in TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", European Union, October 2013, available at http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf, p. 9 to 15.

[93] See also DEPREEUW, S., "De uitzondering voor « tijdelijke technische reproductiehandelingen » na Infopaq I en II en Premier League", *A&M,* 2013, 76-85.

[94] The exception in art. 5(1) InfoSoc Directive has been examined extensively in the final Study of the Study On The Application Of Directive 2001/29/EC On Copyright and Related Rights in The Information Society (The "Infosoc Directive"), p. 109 and f. See also DEPREEUW S., "De uitzondering voor « tijdelijke technische reproductiehandelingen » na Infopaq I en II en Premier League", *A&M,* 2013, 76-85.

[95] ECJ 16 July 2009, Case C-5/08, Infopaq I, par 56.

[96] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par 26

[97] See TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", De Wolf & Partners, October 2013, available at http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf.

[98] ECJ, 4 October 2011, joined Cases C-403/08 and C-429/08, Premier League, par 164

The exception for temporary acts of reproduction is the only mandatory exception provided by the InfoSoc Directive. Consequently, it has been literally implemented in all the Member States of the European Union[99].

Transient or incidental copy. The InfoSoc Directive distinguishes two kinds of temporary acts of reproduction: the transient and the incidental ones.

The transient copy is an ephemeral act of reproduction. According to the decision *Infopaq I*, an act can be held to be 'transient' "*only if its duration is limited to what is necessary for the proper completion of the technological process of which it forms an integral and essential part, being understood that that process must be automated so that it deletes that act automatically, without human intervention, once its function of enabling the completion of such a process has come to an end*"[100]. This means that the transient character of the reproduction should be assessed by reference to the "proper completion" of the technological process (a relative rather than an absolute assessment). The Court does not provide an explicit rule to evaluate how long the copy could last to be qualified as transient.

The word "incidental" means that the reproduction lasts longer than a "transient" copy. The relation between the transient and incidental copies was exposed by the Supreme Court in the case *Meltwater*. The English judge described the role of an incidental copy as follows:

> *"If, as I consider, the copies made in the internet cache or on screen are "transient", it is strictly speaking unnecessary to consider whether they are also "incidental". But I think it clear that they are. The software puts a web-page on screen and into the cache for the purpose of enabling a lawful use of the copyright material, i.e. viewing it. The creation of the copies is wholly incidental to the technological process involved"*[101].

An incidental act of reproduction can be found when the copy is incidental with regard to the main act of exploitation of the work[102], provided that the reproduction may nevertheless not be permanent as it has to remain 'temporary'. That concept of "incidental copy" might prove very useful to legitimate non-"transient" copies required by the use of a technological process to make works available.

Integral and essential part of a technological process. Whether it be transient or incidental, the copy has to be made because it constitutes a step in a technical process of communication[103]. In other words, the copy must enable another use of the work that is executed by means of this technological process. In *Infopaq II*, the Court of Justice held that the fact that the temporary copy initiates or terminates a specific process and the fact that such process involves a human intervention do not alter the conclusion that it may be an integral and essential part of a technological process. The Court interpreted the condition as follows:

> "*The concept of the 'integral and essential part of a technological process' requires the temporary acts of reproduction to be carried out entirely in the context of the implementation of the technological process and, therefore, not to be carried out, fully or partially, outside of such a process. This concept also assumes that the completion of the temporary act of reproduction is necessary, in that the technological process concerned could not function correctly and efficiently without that act.*

[99] Except in The Netherlands, where the Dutch Copyright Act considers that the reproduction right itself does not include "temporary" copies, so that the latter are not part of the exclusive rights, remain from the outset outside the scope of copyright and do therefore not need to be treated as an "exception". In Belgium, the text was implemented with minor deviations in terminology. For more information, see LINDNER, B. and SHAPIRO, T., *Copyright in the Information Society*, Edward Elgar, Cheltenham, 2011, p. 401 (The Netherlands) and 87 (Belgium).

[100] ECJ 16 July 2009, Case C-5/08, Infopaq I, par., 64.

[101] UK Supreme Court, 17 April 2013, [2013] UKSC 18.

[102] TRIAILLE, J.-P., "La question des copies caches et la responsabilité des intermédiaires", in STROWEL, A. & TRIAILLE, J.-P., *Google et les nouveaux services en ligne*, Larcier, 2008, 257.

[103] A. LUCAS, *e.a.*, *op. cit.*, 351.

> *Furthermore, given that Article 5(1) of Directive 2001/29 does not specify at which stage of the technological process the acts of temporary reproduction must be carried out, it cannot be excluded that such an act can initiate or terminate that process.*
>
> *Similarly, there is nothing in that provision to indicate that the technological process must not involve any human intervention and that, in particular, manual activation of that process be precluded, in order to achieve a first temporary reproduction.*"[104]

Transmission in a network between third parties by an intermediary. The temporary copies should "*enable transmission systems to function efficiently, provided that the intermediary does not modify the information and does not interfere with the lawful use of technology, widely recognised and used by industry, to obtain data on the use of the information*" (rec. 33 InfoSoc Dir). This element does not seem to raise many difficulties.

Lawful use. According to the recital 33 of the InfoSoc Directive, a use should be considered lawful if it is authorized by the right holder or if it is not restricted by law. It results that a lawful use may consist of an intended use that is authorized, exempted under a legal exception or one that is not restricted by the applicable legislation.

According to the Court of Justice, the acts of reproduction covered by the exception must not exceed what is necessary for the proper completion of the technological process.

Both in *Premier League* and in *Infopaq II*, the Court of Justice identified the intended purpose of the copy and it assessed whether this was its sole purpose. Then the Court verified whether the intended use was restricted under European or national law, which was not the case. It should be noted that the Court did not require that the intended use be an integral part of the technological process: it suffices that there were no indications that the technical process was used for another purpose.

The Court ruled in *Premier League* that the picking up of the broadcasts and their visual display in private circles does not reveal an act restricted by European Union legislation or by that of the United Kingdom, and concluded that these acts of reproduction have the sole purpose of enabling a 'lawful use' of the works[105]. In *Infopaq II*, the Court of Justice noticed that the technological process used to enable a more efficient drafting of summaries of newspaper articles included several acts of temporary reproduction. It ruled that these acts were not unlawful as the drafting of a summary of newspaper articles is not restricted by the European Union legislation neither by Danish legislation[106]. In *Meltwater*, the UK Supreme Court decided that acts of browsing (including the mere viewing, the access and the consultation of a webpage) constituted a lawful use justifying the making of transient copies generated by an end-user's use of the internet[107].

No independent economic significance. The Court of Justice reminded in *Infopaq II* that the acts of temporary reproduction must facilitate the use of a work or make that use more efficient. The Court admitted that these acts enable the achievement of efficiency gains and, consequently, lead to increased profits or a reduction in production costs. Nevertheless, the economic advantage resulting from these acts of temporary reproduction must not be either distinct or separable from the economic advantage derived from the lawful use of the work concerned and it must not generate an additional economic advantage going beyond that derived from the use of the protected work the technological process concerned[108].

In assessing whether temporary acts of reproduction have independent economic significance within the meaning of Article 5(1) of Directive 2001/29, it is necessary to establish whether an economic advantage

[104] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par. 30-32.

[105] ECJ, 4 October 2011, Joined Cases C-403/08 and C-429/08, Premier League, par. 172.

[106] ECJ, 17 January 2012, C-302/10, *Infopaq II*, par. 42-43.

[107] UK Supreme Court, 17 April 2013, [2013] UKSC 18.

[108] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par. 49-50.

stems directly from the temporary acts of reproduction[109]. Temporary acts of reproduction have an independent economic significance if they generate an additional economic advantage going beyond the advantage derived from the use of the protected work[110]. According to the Court of Justice, there is an independent economic significance if the author of the reproduction is likely to make a profit due to the economic exploitation of the temporary reproduction itself or if the act of temporary reproduction leads to a change in the subject matter reproduced. Such act no longer aims to facilitate the use of the work, but the use of a different subject matter[111].

Reproductions that make access to a work possible have an economic significance (e.g. the display on a television screen)[112]: since the works have an economic significance, access to the works has an economic significance and therefore the reproductions that enable this access have an economic significance. This fact by itself does not preclude the application of the exception, as long as the reproduction does not have an *independent* economic significance. The Court found that the reproductions on a satellite decoder and a television screen are not capable of generating an additional economic advantage, beyond the advantage derived for the intended use (i.e. the mere reception of the broadcast) and that they do not have a separate economic significance. The Court derives this from the fact that these copies are an "inseparable and non-autonomous part of the process of reception"[113].

(ii) Application to data analysis

As we saw above in Part III A, there are many ways to mine a set of data. For the purposes of this Study, we summarized the mining process as follows:

1. The obtaining of the sources;
2. The transformation of the data to fit operational needs;
3. The loading of the data;
4. The analysis of the data, and
5. The drafting of a report.

Those steps are not involved in every mining processes, but they seem rather representative of all the mining techniques which are available for the moment on the market.

In this section, we will detail those five steps and see if the temporary copy exception could apply to them.

1. Obtain the sources

The first part of a mining process is to obtain the sources. Many projects consolidate data from different source systems.

It usually involves **extracting** the data from the source systems but a **direct access** to the sources is also conceivable.

In general, the goal of an **extraction** (the term "extraction" is here not used in the precise meaning of the Database Directive) phase is to convert the data into a single format appropriate for transformation processing.[114] The sources can be created from scratch (*e.g.* a collection of data from public webpages)

[109] Opinion of Advocate General Trestnjak delivered on 12 February 2009, Case C-5/08, Infopaq, par. 127.

[110] ECJ, 4 October 2011, Joined Cases C-403/08 and C-429/08, Premier League, par. 177.

[111] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par. 51-53.

[112] ECJ, 4 October 2011, Joined Cases C-403/08 and C-429/08, Premier League, , par. 174.

[113] ECJ, 4 October 2011, Joined Cases C-403/08 and C-429/08, Premier League,, par. 176.

[114] Wikipedia, http://en.wikipedia.org/wiki/Extract,_transform,_load

or can be provided by a third party (for example, a download from a server or a copy of a "database dump"[115]) or from a publisher.[116]

An intrinsic part of the extraction involves the parsing[117] of extracted data (*i.e.* the analysis of sequences of characters), resulting in a verification that the data meets an expected pattern or structure. If not, the data may be rejected entirely or in part.[118]

There are numerous extraction tools available. Many of them are open source.

When a copy of the data is made from a server to another, it is clear that a reproduction is made. Therefore, the person responsible for the copy should in principle acquire the prior authorization of the holder of the reproduction right.

Temporary reproduction. The question may be asked if the extraction of the sources may fall in the scope of the exception for temporary copies. It should be verified whether the reproduction respects the conditions set out by Article 5.1 of the InfoSoc Directive. We will see that there is no general and clear answer to the condition, since it depends on many factual circumstances.

- **Temporary transient or incidental**. While usually the extraction cannot benefit from the exceptions because if often has a permanent character, some services may be technically based on a reproduction in a cache or RAM memory that lasts for a limited period and is automatically deleted, depending on what the application actually permits or requires in practice. Nevertheless, we saw above that in practice, the national jurisdictions rejected the benefit of the temporary exception to the cache practice of Google, to the technical copies of broadcasts made to enable simulcast and webcast transmissions or the reproduction involved in an online video recorder.
  It is further unlikely that a temporary copy used to mine data is *transient*, the work mostly being available for a certain period of time to be transformed, loaded and/or analyzed. The extraction will not be transient if the removal of the transient copy does not happen automatically and depends on a human intervention.
  It is not excluded that the temporary copy made during the extraction is *incidental* to the main act of exploitation of the work, *i.e.* the analysis of the work.

- **Technological process**. It can be expected that the copy that serves as the basis of the mining process is part of a technological process. In the context of mining processes, the first load of the sources forms the starting point of this process of transmission and it may be qualified as an integral and essential part of those technological processes.

- **Sole purpose.**
  - **Transmission**. The sole purpose of the extraction of the sources does not seem to us to be to enable a transmission in a network between third parties by an intermediary.
  - **Lawful use**. The act of extraction of the sources could fulfill this condition if it has as sole purpose to enable a 'lawful use' of the works. In *Infopaq II*, the process which permits a more efficient drafting of summaries of newspapers articles was not considered unlawful. In *Premier League*, the picking up of broadcasts in private circles was judged as having the sole purpose of enabling a lawful use of the works.

[115] CLARK, J., « Text Mining and Scholarly Publishing », *PRC*, 2012, p.14.

[116] TRUYENS, M. & VAN EECKE P., "Legal aspects of text mining", to be published in *CLSR*, 2014.

[117] As explained in Wikipedia, "parsing" "has slightly different meanings in different branches of linguistics and computer science. Traditional sentence parsing is often performed as a method of understanding the exact meaning of a sentence, sometimes with the aid of devices such as sentence diagrams. It usually emphasizes the importance of grammatical divisions such as subject and predicate. Within computational linguistics the term is used to refer to the formal analysis by computer of a sentence or other string of words into its constituents, resulting in a parse tree showing their syntactic relation to each other, which may also contain semantic and other information".

[118] Wikipedia, http://en.wikipedia.org/wiki/Extract,_transform,_load

If the intended purpose of the mining process is to extract information from various sources and derive new knowledge from them, then it could be qualified as lawful, since information as such, like ideas, is not protected by copyright. Also, the intended use may be authorized by the right holder and hence be considered "lawful". On the contrary, if the right holder expressly reserved his rights in terms and conditions, the extraction of its data will be unlawful.

- **No independent economic significance**. The economic advantage resulting from an act of temporary reproduction must not be distinct or separable from the economic advantage derived from the lawful use of the work concerned and it must not generate an additional economic advantage going beyond that derived from the use of the protected work. It is very difficult to answer this question, because it is relative to the economic value of the data mining process in itself. It seems that every acts involved in the data mining process can have a great economic value. Potentially, we can imagine that the first extraction can have an independent/separate economic significance, but it depends on what the "miner"/"copy-maker" does with the result of the first extraction (*e.g.* if he sells or licenses the results of the extraction). It is thus a question of fact.

  Besides, the acts of temporary reproduction may not lead to a modification of the work.[119] The collection of data does not *a priori* involve a modification of the work.

Another technique is the **direct access** to the data, without any extraction or other download.

This possibility will depend on the infrastructure (technical context) and the other needs of the mining process (*e.g.* a one-time access is sufficient or several accesses during an extended period are needed).

To mine data, a direct access to the content (an online access to a platform, database, collection, materials, individual documents, etc.) on third party servers is technically conceivable. In this case, there is no need to copy the content to a separate location.

The mere access to the *e.g.* online platform is generally not considered as a reproduction (a preliminary ruling is however asked to the ECJ on this question in the case Meltwater – see explanations above about the temporary copy). If there is no reproduction, copyright is not infringed and there is no need to analyze whether the conditions for temporary exception are fulfilled. As indicated however, this will only apply to limited and "simple" TDM projects.

2. Transform

The" transform" stage applies a series of rules or functions to the data extracted from the source in order to derive data for loading into the end target. Some data sources require very little or even no manipulation of data. In other cases, one or more transformation types may be required to meet the business and technical needs of the target database, such as electing only certain columns to load, translating coded values (*e.g.* if the source document stores 1 for male and 2 for female, but the warehouse stores M for male and F for female, the transform stage will standardize the data), encoding free-form values (*e.g*. mapping "Male" to "M" and "Female" to "F"), sorting, joining data from multiple sources and de-duplicating the data, aggregation, etc.[120]

The machine must be able to recognize the data, for example for a database of texts, the words of the sentences of a text, the verbs, and the relationships between the concepts.

[119] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par. 54.

[120] Wikipedia, http://en.wikipedia.org/wiki/Extract,_transform,_load, see the examples to illustrate the transformation types.

During this step, "the text is sliced, diced, chunked and tagged into structured format ready for extraction into a structured database. In essence, this first step deconstructs the human language of text and reconstructs it for machines." [121]

This process requires mostly adaptations (the right of adaptation is not harmonized under the InfoSoc Directive)[122] but, as already explained (see Part III on Exclusive rights), these adaptations may also qualify as reproductions. Some operations may thus be qualified as acts protected under copyright.

Temporary reproduction. The copies that merely serve to transform the works may be exempted under the exception for temporary copy, provided that its conditions are met:

- **Temporary, transient or incidental**. The analysis of the works during a data mining process may involve temporary copies that are kept for the duration of the access to the work. It can involve transient or incidental copies.
- **Technological process**. The copies made during the transformation process are part of a technological process. They may be qualified as an integral and essential part of those technological processes.
- **Sole purpose.** We refer to the considerations above about the extraction of the sources (point 1). In our opinion, the conclusions are identical.
- **No independent economic significance**. We believe that this condition may be fulfilled if the "miner" does not give any separate economic significance to the copies during the transformation phase. However, the transformation phase usually modifies the work (see the examples above in the first paragraph of this point), before the loading step. If it is the case, then the condition is not fulfilled and this step cannot benefit from the exception of Article 5.1 of the InfoSoc Dir., since the ECJ said in *Infopaq II*, that to comply with the condition of not having an independent economic significance, the acts of temporary reproduction cannot lead to a modification of the work.[123]

3. Load

The load phase downloads the data into the end target, it could for example be a server, a hard disk or a data warehouse (which is a central repository of data), also called a corpus[124] or a cube. Depending on the requirements of the organisation, this process varies widely from one project to another. Some data warehouses may overwrite existing information with cumulative information; frequently, updating extracted data is done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals - for example, hourly. Complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.[125]

Temporary reproduction. The exception for temporary copies will not apply to copies resulting from a download. Downloads are in principle permanent reproductions. Indeed, according to the Court of Justice, the temporary copy must result from an automated process that deletes it automatically, without human intervention, once its function of enabling the completion of such a process has come to an end[126]. That condition is not satisfied in the case of a download copy, where it is completely under the control of the "miner"/"copy-maker", hence a human intervention will be necessary to use or erase the copy.

---

[121] CLARK J., « Text Mining and Scholarly Publishing », *PRC*, 2012, p.10 – 11.

[122] TRIAILLE, J.-P., "User Generated Content (UGC) – First Part. Description of the present legal situation regarding copyright in the European Union", *in* TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", European Union, October 2013, available at http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf.

[123] ECJ, 17 January 2012, Case C-302/10, Infopaq II, par. 54.

[124] TRUYENS, M. & VAN EECKE P., "Legal aspects of text mining", to be published in *CLSR*, 2014.

[125] Wikipedia, http://en.wikipedia.org/wiki/Extract,_transform,_load, see the examples to illustrate the transformation types.

[126] ECJ, 16 July 2009, C-5/08, Infopaq I, 64.

However we can imagine a data warehouse of which the content is continuously automatically replaced, so that the storage of the works is not permanent, but could be qualified as temporary.[127] This could be the case on rare occasions. We refer to our analysis of the other conditions above on the collection of the sources (point 1).

The download copy resulting from the use of a download technology will necessarily fall under the reproduction right and it is necessary to obtain the consent of the right holder, except if the download may benefit from another exception (private use, exception for scientific research, see below).

4. Analysis

The next step is to get information from the sources or the data warehouse (if any). This is a multiple step process that can be very complex.[128]

The analysis of the information can consist in many techniques. The application generally asks queries to the data warehouse. The queries are either manually or (more usually) automatically sent to the data warehouse. The applications will often use multiple tools to provide out-of-the box functionalities (comparisons, annotation, recognition of terms or entities, setting of a derivate dataset capable of semantic interrogation, indexing, querying, storage, modification, viewing of searches, categorization, etc.).

These techniques can involve an act of copying. This would not be the case if the copying takes place "word by word" (one word after the other), with no copies of these words being kept. In other cases, if copies are made of copyrighted materials, then a reproduction takes place.

Temporary reproduction. The copies that merely serve to analyse the works (or the derivative works) and to consult them, may be exempted under the exception for temporary copy, provided that the conditions laid down for that exception are met:

- **Temporary transient or incidental**. The analysis of the works during a data mining process may involve temporary copies that are kept for the duration of the access to the work. It can involve transient or incidental copies.
- **Technological process**. The copies made during the analysis are part of a technological process. They may be qualified as an integral and essential part of those technological processes.
- **Sole purpose.** We refer to the considerations above about the extraction of the sources (point 1). In our opinion, the conclusions are identical.
- **No independent economic significance**. We refer to our comment made about the transformation (point 2), which can be applied here.

5. Output

The production of an output is the process of taking the transformed (and analysed) data and presenting it in a way that provides information which is readable by humans. The output (which could also be considered as the data analysis "report") can be presented in many formats, such as a PDF file, an excel file, a written report, a graphic, a website, etc.

At the end of the cycle, the lasts steps may be the archiving and a cleaning up.

Normally, the output does not contain any of the original works that were mined, the works have been analysed and only some information were kept. If it still contains parts of works, for example, if it

[127] TRUYENS, M. & VAN EECKE P., "Legal aspects of text mining", to be published in *CLSR*, 2014.

[128] CLARK J., « Text Mining and Scholarly Publishing », *PRC*, 2012, p.13.

reproduces a short paragraph of a text, then this short reproduction, normally very incidental, could maybe benefit from the quotation exception.

The output is a new work, created by a new author on the base of information contained in previous works. Since no reproduction is made, there is no need to see if the output passes the conditions of the temporary reproduction.

(iii) Preliminary conclusions

During each step of a mining process, we saw that potentially numerous copies are made. If either of those copies is permanent, the copies cannot benefit from the exception of Article 5.1 of the InfoSoc Directive.

In a few rare cases, it is likely that copies involved in the steps of a mining process could fulfil the conditions of the temporary copy; the copies being transient or incidental (for a duration limited to what is necessary for the proper completion of the technological process), being part of a technological process (a mining process), having as sole purpose to enable a lawful use of the work (the only copy of ideas of works, which is not restricted by copyright, or another use allowed by the right holder or another exception) and finally if the copies have no independent economic significance, separable from the economic advantage derived from the lawful use of the work concerned or do not lead to a modification of that work. As one can see, many conditions have to be met.

It means that this exception **will not provide much relief (or really rarely) for data analysis activities**.

b) Exceptions for scientific research under the Infosoc and the Database Directives

We will distinguish between the use of **works** for scientific research in the InfoSoc Directive and the use of a **database** (more exactly, of its selection and/or arrangement) under the Database Directive.

(i) Use of "works" for scientific research – Article 5.3.a) of the Infosoc Directive

1. <u>The principles</u>

In the following pages, we will assess whether data analysis activities could be covered by the exception for scientific research contained in Article 5.3.a) of the Infosoc Directive. Article 5.3 a) provides for an exception to the right of reproduction (Article 2) and to the right of communication to the public (Article 3) when the protected work is used:

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

The exception for scientific research can, in certain circumstances, cover the acts of reproduction performed in the course of data analysis activities, while exceptions to the right of communication to the public are, in our view, not relevant for data analysis. Indeed, as explained in the conclusion in Part III, when the data analysis is made on the basis of data and information which are protected by copyright, the process of data analysis does, in most cases, involve a reproduction of protected materials but no communication to the public nor any making available.

As described in Part III of our Study, the process of data analysis will only in rare cases not involve an act of copying (reproduction), i.e. when the software only "crawls" through the texts or the data, and processes them "one by one", without copying the whole text but only one word or just a few at a time. In such cases, the existence (or not) of possible exceptions to the reproduction right is of no importance, since no infringement is taking place.

Published/unpublished works. It is uncertain whether the exception for scientific research applies only to published works or both to published or unpublished works (we take here the word "published" as meaning that the work has been made accessible to the public by the author or with his consent or that the work has been made public, not in the strict meaning given to the word in the Berne Convention). Article 5 of the Infosoc Directive does not say anything about published or unpublished works and no rule of exclusion of unpublished works from the benefit of copyright exceptions can be found in the *acquis communautaire*. S. Dusollier explains that:

> *"A principle that is not touched upon by the acquis communautaire is whether the exceptions can concern works that have not been published or legally made available. This is a requirement that applies in a number of Member States. In some cases, the exclusion of the benefit of any exceptions applies in a general way and is laid down in an introductory provision to the lists of authorized uses (e.g. Belgium, France, Greece or Italy). In other legislation, the exclusion of unpublished works is a customary condition applied to exceptions (e.g. Germany, Spain) or explicitly appears in the text of specific exceptions (e.g. Denmark, the Netherlands). The terminology also differs, from "published works", "lawfully published or made available", to "works that have been disclosed".*[129]

From a moral rights' perspective, one could argue that a work that has not been made public by his author is a work for which the author has not used his moral right of divulgation. Use of non-divulgated works is in such case an infringement to the moral rights of the author.

This situation is probably not very relevant for TDM, as what has not been published will normally not be available, at least in a licit way, and should thus normally not be available for data analysis either (so that the issue is maybe more theoretical than real). We shall not consider this question here any further.

Not mandatory character of the exception. The exception for scientific research is not mandatory and has been implemented differently in the Member States. Some Member States, amongst those analyzed for the present Study, do not have any exception for research purposes: This is the case for Spain[130] and the Netherlands[131]. Other EU countries (outside of the list of 11 countries which we were asked to analyze in this Study), such as Greece[132] and Slovakia[133], did not implement an exception for scientific research in their national legal order either. In some countries, the exception was not introduced explicitly but case-law has accepted to integrate the exception.

This can be problematic for data analysis activities performed in countries where the exception for scientific research is not in force since Article 5.2 of the Berne Convention[134] refers to the application of

[129] DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses", in TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", European Union, October 2013, available at http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf, p. 253.

[130] Only provided for databases, see Article 34 of the Spanish Consolidated Text of 1 April 1996 on Intellectual Property.

[131] Except for the exception of article 15 of the Dutch Auteurswet allowing for quotation in scientific writings, but this is related to the quotation exception.

[132] LINDNER, B. and SHAPIRO, T., *Copyright in the Information Society*, Edward Elgar, Cheltenham, 2011, p. 241.

[133] LINDNER, B. and SHAPIRO, T., *Copyright in the Information Society,* op. cite note 21, p. 490.

[134] Article 5 - *Rights Guaranteed: […] (2) The enjoyment and the exercise of these rights shall not be subject to any formality; such enjoyment and such exercise shall be independent of the existence of protection in the country of origin of the work. Consequently, apart from the provisions of this Convention, the extent of protection, as well as the means of redress afforded to the author to protect his rights, shall be governed exclusively by the laws of the country where protection is claimed*.

the "laws of the country for which a protection is claimed" (*lex loci protectionis*) to the enjoyment and the exercise of the rights provided under the Convention. That provision is generally presented as a rule of conflict of laws. The rule of Article 5.2 needs to be combined with the principle of territoriality of copyright that commands that the exceptions provided by one national law only applies to the acts of use occurring in its territory. The assessment of whether some use is a copyright infringement or could be excused under an exception thus requires determining first the localization of such use, hence the law under which the use could be confronted with the conditions of the exception for scientific research. As a result, e.g. Spanish law would apply to any claim issued against researchers who are performing data analysis activities in Spain because Spanish laws are the laws for which the protection is claimed. In the absence of a legal provision recognizing an exception for scientific research in the Spanish Copyright Act, the researchers are presumably infringing the copyright of the authors of the works used for data analysis (subject to the application of other exceptions). S. Dusollier states (about the lack of harmonization of the exception in the European Union) that:

> *"The only issue that could be put forward is the absence of any exception in some Member States, which could put researchers in those countries in a less easy situation than their foreign peers."*[135]

Other scenarios can be even more problematic. For example, what would happen if a Spanish research center collaborates with a Belgian university on a project involving data analysis? Data analysis performed in Belgium would fall under the exception for scientific research recognized by the Belgian Copyright Act[136], whereas data analysis performed in Spain would infringe the copyright of the authors of the works used for data analysis in Spain (subject to the application of other exceptions).

The Public Consultation on the review of the EU copyright rules[137], launched by the European Commission on December 5, 2013, specifically identifies this issue as a potential barrier for data analysis[138].

We have not seen this potential problem being illustrated in the literature, but it makes sense to consider that the lack of harmonization of the exception does raise difficulties for users (probably more than for rightholders). This raises the question as to whether the exception should be made mandatory in all Member States – see our recommendations in Part IX of this Study.

In the recent case-law of the European Court of Justice, the European judges have confirmed the principle of strict interpretation of the exceptions contained in the Infosoc Directive. However, this principle has at the same time been somewhat attenuated[139]. The interpretation of the conditions of an exception must "*enable the effectiveness of the exception thereby established to be safeguarded and its purpose to be observed*"[140]. If upheld[141], this case-law would mean that exceptions should not be too literally construed but should be seen as means to the end they pursue".[142]

---

[135] DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses, op. cit., note 4, p 392.

[136] Article 22 of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[137]The Public Consultation was launched by the European Commission on December 5, 2013 and is accessible here: http://ec.europa.eu/internal_market/copyright/initiatives/index_en.htm

[138] See in particular the following questions in the section of the public consultation related to Text and Data Mining: *"(b) [In particular if you are a service provider] Have you experienced obstacles, linked to copyright, when providing services based on text and data mining methods, including across borders?(c) [In particular if you are a right holder] Have you experienced specific problems resulting from the use of text and data mining in relation to copyright protected content, including across borders?*

[139] E.C.J., judgment of 16 July 2009, C-5/08, *Infopaq International*, *ECR*, 2009, I-6569, paragraphs 56-57; ECJ, judgment of 1 December 2011, C-145/10, *Painer*, paragraph 109; ECJ, judgment of 4 October 2011, C-403/08 and C-429/08, *Football Association Premier League and others*, paragraph 162.

[140] ECJ, judgment of 1 December 2011, C-145/10, *Painer*, paragraph 133.

[141] The Painer decision could indeed be interpreted as applying only to the exception of quotation due to its strong relationship with the fundamental freedom of expression.

[142] In this sense, DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses, op. cit., note 4, p 252.

One must thus avoid taking a too narrow construction of the exception for scientific research which would hinder its effectiveness in delivering what it is intended for in Article 5(3)(a) of the Infosoc Directive.

Divergences in implementation. Except for Spain[143] and the Netherlands[144], the exception for scientific research contained in article 5(3) a) of the Infosoc Directive has been transposed in all the Member States that are analyzed by the Study.

These Member States have not transposed this provision in the same way. Some have implemented it in a detailed way; others have enacted national legislation that is more restrictive and narrower than the Directive. In our previous Study[145], S. Dusollier identified and commented on the main differences between the Member States: (a) the diverging objectives pursued by the national provisions, (b) the beneficiaries and the users, (c) the works concerned, (d) the authorized acts, and (e) the implementation of the other conditions contained in Article 5(3) of the Infosoc Directive.

The reason for these differences lies in the optional nature of the exception and the discretion the Member States have in the choice of the measures to transpose the Directive. As outlined in the Green Paper on Copyright of 2008, "*different treatment of the same act in different Member States may lead to legal uncertainty with regard to what is permitted under the exception, especially when teaching and research are carried out within a transnational framework*"[146], whereas the objective was the harmonization of the legislations.

Those differences will be presented briefly in the present section.

a) **Objectives:** The implementation of the exception generally follows the directive and combines in one provision the objectives of teaching and research without treating them differently. Accordingly, research is included in the education exception in Belgium[147], France[148], Hungary[149], Italy[150], Luxemburg[151], and Poland[152]. Generally no specific conditions are defined for research that is subject to the requirements applicable for the teaching objective. Germany separates the purpose of research from the purpose of teaching in two distinct paragraphs but the differences between the two exceptions are minimal.[153]

   Some countries reserve a distinct treatment to uses for research purposes. For instance, Denmark allows the copies of works of art in scientific presentations[154]. This rather restricts the use of copyrighted works for research purposes to one type of activities and one category of works and the limit with the quotation exception seems rather thin[155]. The UK is another example

[143] Only provided for databases, see Article 34 of the Spanish Consolidated Text of 1 April 1996 on Intellectual Property.

[144] Except for the exception of article 15 of the Dutch Auteurswet allowing for quotation in scientific writings, but this is related to the quotation exception.

[145] DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses", in TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", European Union, October 2013, p. 109 et s., available at http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf, pp.352-400.

[146] *Green Paper on Copyright in the Knowledge Economy*, 16.7.2008, COM(2008) 466 final, p. 17.

[147] Article 22, §1, 4*bis*°, 4*ter*° and 4*quater*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[148] Article L122-5, 3°, e) of the French Intellectual Property Code of 1 July 1992.

[149] Article 34, (2) of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

[150] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[151] Article 10, 2° of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[152] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[153] Article 53a of the German Act of 9 September 1965 on Copyright and Related Rights.

[154] Article 23 of the Danish Consolidated Act No. 202 of 27 February 2010 on Copyright.

[155] The exception for quotation is referred to in the article 22 of the Danish Copyright Act with no specific purpose mentioned. The

of a clear separation between education and research[156]. Section 29 of the Copyright, Design and Patent Act admits fair dealing with a literary, dramatic, musical or artistic work for the purposes of research for a non-commercial purpose, whereas educational uses are exempted by another provision, from section 32 to sections 36A. It is worth noting that this provision, as well as the French provision, does not include the term "scientific" used in the Directive to qualify the research, as the British Government considered[157] the insertion of this word as redundant[158].

As explained *supra*, two countries amongst those analyzed for this Study, do not have an exception for research purposes in the sense of Article 5(3)(a) of the Infosoc Directive, Spain[159] and the Netherlands[160].

Interestingly, case law has sometimes had to define the notion of research contained in the exception[161]:

- A UK decision[162] specified that in Article 29 of the Copyright, Designs and Patents Act, which has to be interpreted in the light of Article 5(3) a) of the directive InfoSoc, the purpose of research "must be narrowly interpreted as meaning 'for the purposes of scientific research'"[163] . As regards the condition of illustration required by the European provision, the court held that "it is not clear whether the exception must also be interpreted as being restricted to illustration for … scientific research", which led the judge to assume (rightly so, we think) that the word "illustration" is linked to teaching but not to research[164]. In the same case, the judge decided that the defendants did not reproduce the list of data for the purpose of scientific research but "for the forensic purpose of enabling themselves and others to extract data from mobile phones for use in criminal investigations", which did not amount according to the decision, to scientific research[165].
- Another UK decision insists on the fact that the exception requires not only that the purpose of the act is for research but also that the purpose is a non-commercial one[166]. The conditions are not alternative.

article for inclusion of works of art in scientific presentations just follows, which makes think that the two exception are close.

[156] Article 29 of the British Copyright, Designs and Patents Act of 15 November 1988.

[157] Consultation on UK Implementation of Directive 2001-1929/EC on Copyright and Related Rights in the Information Society: Analysis of Response and Government Conclusions, p. 6.

[158] E. DERCLAYE, "La transposition de la directive 'droit d'auteur dans la société de l'information', *op. cit.,* p.603.

[159] Only provided for databases, see Article 34.

[160] Except for the exception of article 15af the Dutch Auteurswet allowing for quotation in scientific writings, but this could be related to the quotation exception.

[161] W. CORNISH ET AL, *op. cit.,* p. 503, 12.39.

[162] *Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire Police*, High Court of Justice, Chancery Division, 9 November 2011, [2011] EWHC 2892 (Ch), [2012] F.S.R. 15.

[163] *ibidem*, §109.

[164] *ibidem*.

[165] *ibidem*.

[166] *The Controller of Her Majesty's Stationery Office*, Ordnance Survey v Green Amps Limited, Case No: HC07C00249, High Court of Justice, Chancery Division Intellectual Property, 5 November 2007, [2007] EWHC 2755 (Ch), §21 (see the comment of the decision by E. Derclaye, "Of Maps, Crown Copyright, Research and the Environment", *E.I.P.R.*, 2008, p.162. See also T. APLIN, "United Kingdom", in B. LINDNER and T. SHAPIRO (eds.), *Copyright in the information society: a guide to national implementation of the European Directive*, Cheltenham, Edward Elgar, 2011, p. 572.

- A decision of the Danish Eastern Court of Appeal, in 2000, considered that an art book/journal published a couple of times a year by the art museum 'Louisiana' was not for scientific research[167].
  In another decision[168], the same court decided that a newspaper feature article and a newspaper review were not equivalent to scientific research.

b) **Beneficiaries and users**: Beneficiaries of the research exception are not defined in the current exception that adopts a functional approach based on the activity of doing research. As a result, anyone could claim the benefit of the exception, irrespective of her profession, training, or relation to a recognized research institute. Sometimes, Member States have stated that the benefit of the exception concerns also research establishments, or else particular individuals. Some domestic laws have even been very specific as to who may or may not rely on the exception. On the contrary, others followed the wording of the Directive. This lack of determined beneficiaries in the drafting of first, the Directive, and then in the national transposition laws, can be understood as an attempt to apply the exception to all research institutions, which is confirmed by the texts of recital 34[169] and recital 42[170] of the Directive; the former speaking about "public institutions" and the latter referring to the notion of "establishments concerned".

   States that have not included in their domestic law a specific beneficiary include Hungary, Denmark, Italy, Luxemburg, and the UK.

   Other Member States restrict the benefit of the exception to certain beneficiaries and/or users: in Poland, the exception applies to research and educational institutions (so, not to individual researchers)[171]. Belgium limits the making available through the closed networks of the concerned institutions to 'the establishments recognized or officially organized for that purpose by the public authorities'[172]. France allows "uses of short works or excerpts for purposes of illustration or analysis, but only if these works or excerpts are communicated within strictly limited circles of students, teachers, or researchers"[173]. In Germany, the exception authorizes the reproduction of protected works "for one's own scientific use" and the making available "exclusively for a specifically limited circle of persons for their personal scientific research to be made available to the public".[174]

   Divergences in implementation of the Directive on this issue could however raise difficulties for data mining activities (allowed for some beneficiaries in certain Member States under this exception and not in others).

c) **Works concerned**: The scope of the national implementation of Article 5(3)(a) of the Infosoc Directive is also sometimes limited to a determined subject matter. Indeed, some Member States tend to limit the "use" provided under the exception for scientific research in Article 5(3)(a) of the Infosoc Directive to "extracts of works" in their national legal order:

[167] Ugeskrift for Retsvaesen 2000.1291Ø.

[168] Eastern Court of Appeal, 17 May 2002.

[169] "Member States should be given the option of providing for certain exceptions or limitations for cases such as educational and scientific purposes, for the benefit of public institutions such as libraries and archives, for purposes of news reporting, for quotations, for use by people with disabilities, for public security uses and for uses in administrative and judicial proceedings".

[170] "When applying the exception or limitation for non-commercial 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".

[171] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights, *op. cit.,* p. 301.

[172] Article 22, §1, 4*quater*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[173] B. LINDNER and T. SHAPIRO (eds.), *Copyright in the information society: a guide to national implementation of the European Directive*, Cheltenham, Edward Elgar, 2011, p. 204.

[174] Articles 52a and 53 of the German Act of 9 September 1965 on Copyright and Related Rights.

- In Belgium: “articles or works of fine art in part or in whole or short fragments of other works”;[175]
- In the UK: “literary, dramatic, musical or artistic work”[176]
- In Denmark: “works of art and works of a descriptive nature for the use in scientific presentation”;[177]
- In France: “extracts of works” ;[178]
- In Germany: “small parts of a work, small-scale works or individual articles released in newspapers or periodicals or made available to the public” ;[179]
- In Italy: “fragments or parts of works” ;[180]
- In Poland: “fragments of disseminated works” ;[181]
- In Hungary: “part of a literary or musical work or such a type of an entire work of a minor volume disclosed to the public”;[182]
- In Luxembourg: “short fragments of works” (in the case of publication on the internet).[183]

Moreover, in Germany[184], copies in graphic data file are permitted in quantities required for the person receiving instruction (which indicates that this applies more to education/instruction than to research). Concerning the provision 52a UrhG, the Court of Appeal of Stuttgart of the 4 April 2012 has held that an extract of 91 pages of a psychology textbook made available to students by a teacher on an electronic platform is not a “small part” of the textbook[185]. This condition must be assessed on a case-by-case basis, from a quantitative point of view and in relation to the work as a whole and its content[186]. The Court sets a limit of 3 pages beyond which the exception can no longer be claimed, for purposes of legal certainty. This long enumeration reveals that Member States tend to limit the use to extracts of works. This can be problematic for data analysis which involves, most of the time, the use of the works in full (see our analysis *infra*).

d) **Authorized acts**: All authorized acts do not apply to all works covered by the exception in the domestic laws for there are large variations among Member States.

The act authorized by the Directive is included under the general term of "use". Article 5(3) of the Directive expressly refers to Articles 3 and 4, which deal respectively with the reproduction and communication to the public. In addition, Article 5(4) states that when the Member States may provide for an exception or limitation to the reproduction right under paragraph 3, they may also provide for an exception or limitation to the right of distribution covered by Article 4, to the extent justified by the purpose of the authorized act of reproduction. Such use may be interpreted as exempting any type of acts of exploitation, therefore including making available online and digitization. In majority though, the other Member States have decided to specify the authorized acts.

[175] Article 22, §1, 4*bis*° and 4*ter*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights. A royal decree is expected to give a new formulation of Article 22, §1, 4*bis*° including work of graphic art.

[176] Article 29 of the British Copyright, Designs and Patents Act of 15 November 1988.

[177] Article 23 of the Danish Consolidated Act No. 202 of 27 February 2010 on Copyright

[178] Article L122-5, 3°, e) of the French Intellectual Property Code of 1 July 1992.

[179] Articles 52a and 53 of the German Act of 9 September 1965 on Copyright and Related Rights.

[180] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[181] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[182] Articles 34 (2) of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

[183] Article 10, 2° of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[184] Article 53 of the German Act of 9 September 1965 on Copyright and Related Rights.

[185] OLG Stuttgart, 04.04.2012, 4 U 171/11, GRUR 2012, 718.

[186] *Propriété Intellectuelle,* octobre 2012, n° 45, p. 487.

First of all, **reproduction** is encompassed in all the national provisions, to a greater or lesser extent. Most of the time, this act of reproduction permits the digital reproduction, explicitly or not. For instance, in Belgium, the law draws a distinction between the reproduction on paper or any similar medium and the reproduction in any medium other than paper or similar medium, the latter aiming clearly at the reproductions under a digital format.

Communication is referred to in Belgium[187], France[188], Luxemburg[189], and Italy[190], along with publication in Italy[191], making available to the public and transmission in Germany[192], borrowing in Hungary[193], summary and quotation as well as publication on the internet in Italy[194], use in Poland[195].

Some domestic laws also mention **translation**: for instance, in Poland[196].

a) **Other conditions** : In compliance with Article 5(3) a) of the Directive, the **non-commercial nature** of the purpose pursued is implemented in Belgium[197], Denmark[198], France[199], Germany[200], Hungary[201], Italy[202], and United Kingdom[203] , but not in Luxembourg[204] and Poland[205].

A UK decision stresses that the fact that an establishment is publicly funded "is not determinative of whether the use was a non-commercial purpose"[206]. In that specific case, the judge considered that the purpose of law enforcement (for use in criminal investigations) was not a non-commercial purpose[207]. Moreover, according to the judge, the wording of the European provision related to

[187] Article 22, §1, 4*quater*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[188] Article L122-5, 3°, e) of the French Intellectual Property Code of 1 July 1992.

[189] Article 10, 2° of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[190] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[191] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[192] Article 52a and 53 of the German Act of 9 September 1965 on Copyright and Related Rights.

[193] Article 33, (4) of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

[194] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[195] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[196] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[197] Article 22, §1, 4*bis*°, 4*ter*° and 4*quater*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[198] Article 23 of the Danish Consolidated Act No. 202 of 27 February 2010 on Copyright.

[199] Article L122-5, 3°, e) of the French Intellectual Property Code of 1 July 1992.

[200] Articles 52a, 53 and 53a of the German Act of 9 September 1965 on Copyright and Related Rights.

[201] Article 33, (4) of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

[202] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[203] Article 29 of the British Copyright, Designs and Patents Act of 15 November 1988.

[204] Article 10, 2° of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[205] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[206] *Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire Police*, High Court of Justice, Chancery Division, 9 November 2011, [2011] EWHC 2892 (Ch), [2012] F.S.R. 15, §110.

[207] *ibidem*.

"the extent justified by the noncommercial purpose to be achieved" has to be understood as the British notion of fair dealing[208]. To appreciate the notion of fair dealing, the judge refers to three most important factors laid down in the decision *Ashdown v Telegraph Group Ltd*[209] : "(1) the degree to which the alleged infringing use competes with exploitation of the copyright work by the owner (…); (2) whether the work has been published or not…. ; (3) the extent of the use and the importance of what has been taken (…)"[210]. In the decision mentioned, the defendants' reproduction of the list of data competed with the exploitation of works by the owner, the list was unpublished and the extent of the reproduction was considerable, three factors that weighted against of fair dealing[211]. In addition, another UK decision specifies the moment when the commercial or non-commercial nature of the research has to be taken into account: "Presumably, any research which, at the time it is conducted, contemplated or intended, should be ultimately used for a purpose which has some commercial value will not be within the permitted act"[212].

Acknowledgment i.e. the indication of the source including the author's name is sometimes mentioned in the legal provision itself, which transposes the exception, but also sometimes provided for in a separate article, like in Denmark[213]. It is not required in Poland[214], but it is in Belgium[215], France[216], Germany[217], Luxemburg[218], Hungary[219], Italy[220], Denmark[221], and in the UK[222], the latter requiring a sufficient acknowledgement[223] – most of those countries providing for this indication "except when this is impossible" (as the Directive allows).

The Directive does not provide for any **compensation or remuneration** for the research exception, but allows it under recital 34.

This diverging implementation of the terms of Article 5(3)(a) constitutes a barrier for data analysis activities. As a matter of fact, data analysis made for scientific research generally implies the use of the work in full because otherwise, in many cases, it would fail to fulfill one of its primary purposes which is to rely on an exhaustive and accurate *corpus* of works – which is certainly necessary to qualify as "scientific" (*cf. infra*). It may be that the use of works in full is not always a prerequisite for data analysis but the exclusion of such use would hinder the development and the expansion of many data analysis activities.

[208] *ibidem*.

[209] *Ashdown v Telegraph Group Ltd [2001] EWCA Civ 1142; [2002] Ch.149.*

[210] *Forensic Telecommunications Services*, *op. cit.,* §111.

[211] *ibidem*., §112.

[212] *The Controller of Her Majesty's Stationery Office*, Ordnance Survey v Green Amps Limited, Case No: HC07C00249, High Court of Justice, Chancery Division Intellectual Property, 5 November 2007, [2007] EWHC 2755 (Ch), §23. The decision refers to Copinger, § 9-28.

[213] Article 11 of the Danish Consolidated Act No. 202 of 27 February 2010 on Copyright.

[214] Article 27 of the Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[215] Article 22, §1, 4*bis*°, 4*ter*° and 4*quater*° of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[216] Article L122-5, 3°, e) of the French Intellectual Property Code of 1 July 1992.

[217] Article 63, §2.

[218] Article 10, 2° of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[219] Article 33, (4) of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

[220] Article 70 of the Italian Law No. 633 of 22 April 1941 for the Protection of the Copyright and the Other Related Rights to its Exercise.

[221] Article 11 of the Danish Consolidated Act No. 202 of 27 February 2010 on Copyright.

[222] Article 29 of the British Copyright, Designs and Patents Act of 15 November 1988.

[223] Definition in Section 178.

As it stands, the text of Article 5(3)(a) of the Infosoc Directive allows acts of reproduction, communication to the public and distribution consisting in the use of protected works for scientific research. There are no indications, in the Articles or in the Recitals of the Infosoc Directive that the "use" should be limited to "extracts of the protected works". Member States are free to further limit the scope of the facultative exceptions of the Directive (and to implement them or not), except if, based on recent case-law of the ECJ one should come to the conclusion that such implementation of the exception limiting it to extracts of works would be contrary to the exception's purpose. Quite often, this limitation may be found in legislation because one and the same provision deals with "illustration for teaching" and with "scientific research". It seems logical that only parts of works are needed to "illustrate" an educational work; this seems much less logical for scientific research activities.

Cumulative conditions under Article 5.3.a) of the Infosoc Directive. The use of works (data, texts, images, videos, etc.) for the purpose of data analysis will not infringe the authors' exclusive rights if the user can prove that (cumulative conditions):

(i) works are used for the sole purpose of illustration for teaching or scientific research;

(ii) the source, including the author's name, is indicated, unless this turns out to be impossible;

(iii) works are used to the extent justified by the non-commercial purpose to be achieved.

Furthermore, the exception can only be applied to the extent that (iv) the use of the works does not conflict with a normal exploitation of the work or other subject-matter and does not unreasonably prejudice the legitimate interests of the rightholder (three-step test).

Distinction between "scientific research" and "non commercial purposes". We do not think the requirements "scientific research" and "non commercial purposes" are redundant, for the following reasons:

- Scientific research may be made for commercial purposes or for non-commercial purposes:

  a. "*Fundamental*" scientific research by a university, not aiming at commercially exploitable results and whose results will be made publicly available is made for non-commercial purposes;
  b. "*Applied*" scientific research made by a private company, like a pharmaceutical company, aiming at developing new drugs, would certainly qualify as scientific research but is aimed at putting a new drug on the market and then (legitimately) deriving profits from it; the results will usually be patented or – if not patentable – be kept confidential: such scientific research is made for commercial purposes;

- Data mining for commercial purposes (or even for non-commercial purposes) may sometimes qualify as "scientific research":

  a. this would be the case if the purpose of the data mining is of a scientific research nature;
  b. and this would not be the case if the data mining does not contribute to scientific research but is either meant for e.g. marketing purposes (commercial purposes).

This, we think, clearly illustrates that **scientific research and non-commercial purposes are two separate conditions** Also, it has been said that borderline cases will unavoidably exist.

2. Application to data analysis

We will now apply these conditions to data analysis:

a. The works are used for the sole purpose of scientific research

Scientific research. Article 5.3.a) provides an exception to copyright when protected works are used for "scientific research". What exactly does the term "scientific research" mean"?

According to the Explanatory Memorandum, paragraph 36: "*the term 'scientific research` within the meaning of this Directive covers both the natural sciences and the human sciences*".

M. Walter and S. Von Lewinski explain that:

> *"While research is the exploration of a certain subject matter in order to find data or any other kind of information or to gain knowledge, "scientific" research must be carried out in a methodical and systematic way".*[224]

S. Dusollier adds that:

> *"Some domestic laws, when implementing the directive, have dropped the adjective 'scientific', on the basis that research is necessarily scientific. This was the case in the UK where this argument was explicit.*[225]

However, UK Courts have recently ruled that the purpose of research must be narrowly interpreted as meaning 'for the purposes of scientific research'"[226].

In our view, not all research is necessarily "scientific". Although it is *necessary* for a research to be carried out in a methodical and systematic way to be qualified as scientific, it is clear that it is not a *sufficient* condition. A research can be methodical and systematic yet not be scientific (this is not true the other way around). For example, predictive marketing research (which will generally qualify as "commercial research")[227] may be carried out in a methodical and systematic way, but does not in our view qualify as scientific research. Similarly, when a person re-makes a research already undertaken by others, on the basis of the same assumptions and by using the same data and ending up with the same conclusions, by using methodical and systematic methods, this cannot qualify as scientific research.

Since the word "scientific" was attached by the legislator to the word "research" in the InfoSoc Directive, one should normally conclude that the legislator intended to qualify the word "research" and that a distinction is to be made between "research" and "scientific research". We could not find significant case-law or comments on the "scientific" condition. Judges would have to assess this condition on a case-by-case basis, but in our opinion (and recognizing that this would need to be confirmed by case-law), "scientific" refers probably to the fact that a "scientific" research adds something to the state of science (to its knowledge at a particular moment), whether it be to confirm or to infirm a theoretical hypothesis.

The purpose of data analysis is often described as being to uncover new insights from previously known data. It must be checked, on a case-by-case basis, whether a given data analysis project adds something to the state of science in order to be qualified as scientific research.

This raises, from a policy standpoint, the question as to whether the exception should benefit research in general, whether it is scientific or not; however, only scientific research is taken into consideration by the

[224] WALTER, M. W., and VON LEWINSKI, S. V., *European Copyright,* op. cit. note 1, p. 1043.

[225] "A number of right owner organisations suggested that s.29 [of the UK Copyright, Design and Patent Act 1988 on Exception for research and private study] should also refer to 'scientific' research as does Article 5.3(a) [of the Infosoc Directive], but the Government thinks this unnecessary since the breadth of this term is such that it would not appear to add anything meaningful", *Consultation on UK Implementation of Directive 2001/29/EC on Copyright and Related Rights in the Information Society: Analysis of Responses and Government Conclusions*, October 2002, p.6.

[226] *Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire Police*, High Court of Justice, Chancery Division, 9 November 2011, [2011] EWHC 2892 (Ch), [2012] F.S.R. 15, § 109.

[227] Therefore, the distinction regarding marketing will mostly be whether it is for commercial purposes or not (the answer being affirmative).

InfoSoc Directive. This still leaves room for scientific data analysis projects, which means that the exception is useful and relevant for TDM.

Illustration. The reference to the term "illustration" in the exception for scientific research is controversial and differs from one Member State to another. The question has been raised, on the basis of the ambiguous drafting of Article 5.3.(a), as to whether the exception only allows "illustration for scientific research" or more broadly "scientific research". "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 or a comparison".[228] In a research context, "illustration" could be understood as allowing the researcher to reproduce or otherwise use a work (or arguably only part of it) "as an example". This is a critical requirement with regard to the applicability of the exception to data analysis. Some Members States do not refer to "illustration" at all; some others use it in connection with "teaching" and not with "scientific research". S. Dusollier (like other authors[229]) recently argued that:

> *"Contrary to teaching purposes, the requirement of "illustration" seems not to apply to scientific research. This is sometimes considered as uncertain by commentators, as the text might receive different interpretation depending on the language version".*[230]

This interpretation is in line with the recent case-law in the UK [231].

One could in our view sustain that the purpose of the exception was not to be limited to the "illustration for scientific research", because this would only cover very limited uses and cases and would be of very limited use (quotation in research reports only?). If the exception must be interpreted in a manner which meets its purpose, our view is that it should not be limited to illustration for scientific research.

Would the exception be interpreted as allowing the use of works solely for the purpose of "illustration" for scientific research, then data analysis on protected works would be – in all and every cases - excluded from the benefit of the exception accordingly. Data analysis involves *de facto* the use of the work for enriching the scientific research taking place, and not just for "illustration" of scientific research: a great number of works are usually copied (and copied in full) in a normal TDM project: should the use be limited to a use for "illustration", most projects would become impossible to realize.

Sole purpose. Scientific research must be the "sole purpose" of the use for which the exclusive rights may be restricted. M. Walter and S. Von Lewinski explain that (although they refer to the exception for the purpose of illustration of teaching, the reasoning is the same here):

> *"When a work is reproduced in a book that is not exclusively designated for teaching purposes at schools and similar institutions but is a general non-fiction book that explains a certain subject*

[228] DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses, op. cit., note 4, p 375.

[229] "The revised provision could directly permit —use solely for the purposes of teaching or scientific research and thus remove any reference to the confusing term for —purpose[s] of illustration". REICHMAN, J.H. and OKEDIJI, R.L., "When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale", 2012, Minn. L. Rev., *I.I.C.*, vol. 43, 2012, p. 1432.

[230] "The English text (purpose of illustration for teaching or scientific research) grammatically indicates that illustration applies only to teaching, by want of the preposition 'for' before 'scientific research'. It is even more clearer in the German text that says: 'für die Nutzung ausschließlich zur Veranschaulichung im Unterricht oder für Zwecke der wissenschaftlichen Forschung', clearly separating the two purposes. On the other hand, the French ('à des fins exclusives d'illustration dans le cadre de l'enseignement ou de la recherche scientifique') or Italian ('finalità illustrativa per uso didattico o di ricerca scientifica') versions sound as relating illustration to both teaching and research. In French, in order to ascertain that illustration would not apply to research, it would have been better to formulate the exception as follows: à des fins exclusives d'illustration dans le cadre de l'enseignement ou de recherche scientifique, by deleting the preposition 'la'", DUSOLLIER, S., "*The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses, op. cit.,* note 4, p 378.

[231] *Forensic Telecommunications Services Ltd v Chief Constable of West Yorkshire Police*, High Court of Justice, Chancery Division, 9 November 2011, [2011] EWHC 2892 (Ch), [2012] F.S.R. 15, § 109.

> *matter and thus addresses the general public, but may also be used for teaching, such reproduction would not be privileged by paragraph (3)(a).*"[232]

This condition imposes to pursue one purpose: scientific research. One must be careful to distinguish this condition from the third condition of Article 5(3)(a) of the Infosoc Directive. Indeed, the third condition provides that the work can be used "to the extent justified by the non-commercial purpose to be achieved" (i.e. the use must be necessary in order to achieve the non-commercial purpose). This means that scientific research must be *the* ***sole purpose*** behind the use of the protected works for data analysis. Data analysis would fall outside of the scope of the exception from the moment it is done also for other purposes that scientific research, i.e statistical, behavioural analysis (which do not qualify as scientific research), etc[233]. This condition may constitute a barrier for some data analysis projects and one may wonder why the existence of other purposes (even if non commercial) should *ipso facto* render the exception not applicable. This raises the question as to whether the "sole purpose" condition should not be better replaced by a "**main purpose**" condition; we will come back to this when discussing recommendations – yet the observation arguably concerns the exception for scientific research in general, and not just its application or relevance for TDM.

In conclusion, if one accepts that the exception does not only apply "to illustrate" scientific research (in which case it would become more or less useless for data analysis); the "scientific research" condition is useful for data analysis. It does exclude "non scientific" research or data mining projects which do not even qualify as "research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research projects" but this does not mean that the distinction is not adequate.

The manner in which it has been implemented may cause difficulties, particularly if it is limited to "extracts of works" or only benefits certain kinds of beneficiaries. This illustrates a lack of harmonization and raises the question of its mandatory character (which we will discuss later).

b. <u>The source, including the author's name, is indicated, unless this turns out to be impossible</u>

The second condition of Article 5(3)(a) of the Infosoc Directive imposes the indication of the source, including the author's name.[234] However, the user is only obliged to indicate the source, including the author's name, provided that this "does not turn out to be impossible" (*impossibilium nulla est obligatio*). M. Walter and S. Von Lewinski explain that:

> *"There may be cases of legal impossibility, in particular, where the author has chosen to stay anonymous and the mentioning of his name, if known to the user, would even violate his moral rights. It may be more difficult to determine when such obligation is in fact (factually) impossible; the Directive does not indicate what efforts must be made to find the author's name or other indication of source before such indication may be considered impossible".*[235]

Regard the efforts which the researcher should make, maybe a kind of "diligent search" criterion would be a reasonable compromise.

---

[232] WALTER, M. W., and VON LEWINSKI, S. V., *European Copyright,* op. cit. note 1, p. 1044.

[233] It undeniable that the commercial/non-commercial distinction could also help in eliminating certain projects from the benefit of the exception, yet the "scientific research" condition" and the "non-commercial purpose" condition do not fully overlap, be it simply because there are obvious "commercial scientific research projects" (by the pharmaceutical sector, just to give one example).

[234] "Beyond the author's name, the source includes the title of the work and the publishing house or another place, including a website, from which the work or other subject matter has been taken". WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1044.

[235] WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1044.

Data analysis is basically associated with quantity. Data analysis may involve the processing of hundreds or thousands of works from different sources. An obligation to carry out a diligent search and to indicate the source, including the author's name, of each individual work, although it might be technically feasible, could sometimes constitute a serious – even insurmountable – barrier for the market actors.

One might also **question the relevance of this obligation** in the context of data analysis for the following reason:

- this requirement seems logical for the exception regarding illustration for teaching because in that case, a work is being reproduced (or communicated to the public) and is thus "visible" in the teaching materials (so that every copy of such materials involve a copy of the pre-existing works): it is thus understandable that, in such cases (like for the quotation exception) the source, including the author's name, must be indicated;
- in many scientific research projects however (and particularly in TDM projects), the research project or process may necessitate to copy pre-existing works but it may well be (and, in TDM projects, it will almost always be the case) that the research output does not include any part of the pre-existing works but only draws conclusions (new insights, new patterns) thereon (see indeed our conclusion to the first part of this Study and the difference we make between the data analysis process and the data analysis output). In such situations, a copy of the output (as opposed to a copy of the teaching materials) will not involve any copy of the preexisting works. One may question whether it is still justified to impose that the author's names of all preexisting works used for (i.e. copied during the process of) the research be mentioned[236]; maybe the requirement was introduced because both the educational exception and the research exception were dealt with in one single provision? Furthermore, in TDM cases, the list of "sources" risk being very long: what will be the particular interests and benefits the author of being mentioned (one could say, "lost") in a long list of hundreds of other authors? And one may wonder what would be the damage of not being mentioned amongst hundreds of other authors.

However, the obligation to mention the source is not absolute. Article 5(3)(a) provides for an important safeguard in case of impossibility to indicate the sources of the works. This escape clause could be useful for data analysis since in case of impossibility to mention the source, the work can still be used and inserted in the corpus to be mined.

Therefore, **in conclusion on the requirement to mention the source or the name**, the second condition of Article 5(3)(a) – as it stands – does contain significant flexibility (safeguard clause in case of impossibility) to allow data analysis. Although it constitutes a burden, it cannot be considered as such as a barrier to data analysis. One may however wonder if authors actually benefit from this burden[237].

c. <u>The works are used to the extent justified by the non-commercial purpose to be achieved</u>

Only acts accomplished with a non-commercial purpose and justified by the non-commercial purpose will benefit from the exception in Article 5(3)(a). It is commonly admitted that "commercial" should be read as

[236] This is all the more true that, for ethical or reputational considerations, the researchers will usually spontaneously cite their sources – be it to avoid being criticised by their peers.

[237] In the UK draft legislation on TDM, the following is provided (new Section 29A and a new Annex 2C on "Data analysis for non-commercial research"):

> (1) "*Where a person has lawful access to a copy of a copyright work, copyright is not infringed where that person makes a copy of the work for the purposes of carrying out an electronic analysis of anything recorded in the work provided that:*
>
> *(…)*
>
> *(b) the copy is accompanied by sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)".*

including direct and indirect economic and commercial advantages[238]. However, this criterion is very difficult to apply in practice and is full of uncertainties.

The criterion "commercial v. non-commercial" is not often used in copyright law. We can think of two other copyright cases where this criterion is relevant. First, the Directive on rental right and lending right[239] refers to the distinction between commercial and non-commercial to distinguish lending from rental. This distinction applies to the act of lending and not to the beneficiary. Article 1 gives a definition of these terms:

> "*Rental" means making available for use, for a limited period of time and for direct or indirect economic or commercial advantage;*
>
> *"Lending" means making available for use, for a limited period of time and not for direct or indirect economic or commercial advantage, when it is made through establishments which are accessible to the public.*

However, the payment of a fee for lending does not exclude the qualification of lending. As explained in Recital 11 of this Directive:

> "*Where lending by an establishment accessible to the public gives rise to a payment, the amount of which does not go beyond what is necessary to cover the operating costs of the establishment, there is no direct or indirect economic or commercial advantage within the meaning of this Directive".*

Secondly, the Infosoc Directive provides for an exception to the exclusive right of reproduction for certain non-profit making establishments, such as publicly accessible libraries and equivalent institutions, as well as archives. Article 5(2)(c) of the Infosoc Directive provides for a facultative exception to the reproduction right provided for in Article 2:

> *"c) 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".*

The specific acts must not be "for direct or indirect economic or commercial advantage". In the text of article 5(2)(c), this factor can be understood as referring both to the eligible beneficiaries and to the nature of the authorized acts. The wording "for direct or indirect economic or commercial advantage" is identical to the ones used in the Directive on rental and lending rights. One may thus say that an act not carried out for direct or indirect economic or commercial advantage must 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". But there is – strangely enough - not much more indication on this distinction to be found in other directives, in case-law or in the literature.

This solution is tailored to non-profit making establishments. However, Article 5(3)(a) does not limit the beneficiaries of the exception for scientific research to non-profit making establishments.

Recital 42 of the Infosoc Directive, which relates specifically to the teaching and research exception, provides useful guidance on that matter:

> *"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 organizational structure and the means of funding of the establishment concerned are not the decisive factors in this respect".*

[238] WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1045.

[239] Directive 2006/115/EC of the European Parliament and of The Council of 12 December 2006 on rental right and lending right and on certain rights related to copyright in the field of intellectual property (codified version), *JO* L 376/28, 27 December 2006.

One might wonder what the decisive factors are in relation to "the activity as such". In addition to criteria that are obvious or often quoted (university v. pharmaceutical company), we could think of several facts or factors which may help identifying the non-commercial nature of the activity:

- if the funding agreement between the funder and the funded research institution/center provides that the research institution/center will own the intellectual property rights on the results, it means that the funder is not put in a position where it can later on commercially exploit the results of the research (except under a license from the research institution/center), one can argue that the research is then also not commercial (the only objection could be if, for the research institution, the expected exploitation of the results have a commercial character); in such case indeed, even if the funding comes from commercial sources, the activity for the funder has no commercial character since he will not be able to exploit the results; we would consider this as a non-commercial character of the research;

- the same would hold true if the funding agreement between the funder and the funded research institution/center provides that the research results are not confidential and may be made public and/or that the research output will be published in open access (or other open patent or similar formulas); in such case, the criterion to decide whether the activity is commercial or not does not relate to the funding but to the possible future exploitation of the results. If they are available to everyone, they are not monopolized by one entity and are not used in any other manner to generate revenues (e.g. advertisement based); this would look like a non-commercial purpose also: the institution doing the research has no guarantee whatsoever that it will enjoy the ownership of the IP rights on the results;
- if a private company finances research for philanthropic purposes (e.g. in a sector in which it is not even active, like a telecom company funding AIDS research), this should be considered as non-commercial (the only possible objection could be that the company is looking for an indirect economic – reputational – advantage but this objection could in many cases be set aside);
- some research projects have as their purpose to bring a new product on the market; a priori, this would qualify as research with a commercial purpose. Some other research, typically "fundamental research" does not have such objective; a priori, this would qualify as research with no commercial purpose;
- 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 of the word – which then prevents anyone from patenting research on that basis; we believe that typically, this would also qualify as an example of research without a commercial purpose.

It has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion. We will discuss this further when making recommendations.

It is worth noting that the non-commercial criterion has been the object of in-depth research by the Creative Commons (CC). In 2009, the Creative Commons published a "Study of How the Online Population Understands "Noncommercial Use"?". As it stands, the non-commercial criterion (NC) used in the various CC licenses is the following:

> *"Non-Commercial means not primarily intended for or directed towards commercial advantage or monetary compensation".*

d. <u>No conflict with a normal exploitation and no unreasonable prejudice to the legitimate interests of the rightholder (three-step test)</u>

Once all the conditions of Article 5(3)(a) are fulfilled, the exception must be conforming to the "three-step test" provided for in Article 5(5) of the Infosoc Directive.

Article 5(5) of the Infosoc Directive provides that:

> *"5. The exceptions and limitations provided for in paragraphs 1, 2, 3 and 4 shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder".*

Accordingly, the exception for scientific research for data analysis is only permitted:

(i) in certain special cases;

(ii) which do not result in a conflict with the normal exploitation of a work and;

(iii) which do not unreasonably prejudice the legitimate interests of the author (or other right-holder).

The interpretation and application of the three-step test can be difficult. There is no consensus to date as to who are the addressees of the "three-step test": the Member States and/or the national courts?

The interpretation of the "three-step step" has been the object of discussion at the European and at the International levels. At the European level, the Court of Justice of the European Union:

> *"has so far refrained from interpreting the criteria of the triple test and has considered, in two cases related to the temporary copying exception, that the compliance with the conditions of that exception suffices to satisfy the three-step test*[240]*".*[241]

We shall not discuss here whether Article 5.3.(a) as it is drafted does satisfy the three-step test. Only if changes are suggested to it for data analysis purposes – or if a new specific TDM exception is considered, should these suggested changes be examined on their conformity with the test. In this respect, the possible adjunction of a system of fair compensation may be useful in ensuring such conformity.

We will discuss this later when making recommendations.

In this regard, M. Walters and S. Von Lewinski explain that:

> *"[…] The Directive does not require that the right holders receive fair compensation. However, the application of the three-step test under Article 5(5) of the Infosoc Directive may result in an obligation of the Member States to provide for some form of fair compensation or remuneration, depending on the individual exception or limitation under national law.*[242]

Systems of fair compensation do necessitate the intervention of collective mechanisms and raise difficult cross-border issues.

e. Preliminary conclusions

As preliminary conclusions on the exception for scientific research under the Infosoc Directive:

- if one accepts that the exception does not only apply "to illustrate" scientific research (in which case it would become more or less useless for data analysis), the "scientific research" condition is useful

[240] ECJ, judgment of 4 October 2011, C-403/08 and C-429/08, *Football Association Premier League and others*, paragraph 181; ECJ, order of 17 January 2012, C-302/10, *Infopaq International* (Infopaq II), paragraph 56.

[241] DUSOLLIER, S., "The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses, op. cit., note 4, p 256.

[242] "As an example of such remuneration in the case of reproduction of works for textbooks, see § 46(1) and (4) of the German Copyright Act, which already existed before the adoption of the Infosoc Directive". WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1045.

for data analysis. It does exclude "non scientific" research or data mining projects which do not even qualify as "research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research projects" but this does not mean that the distinction is not adequate;

- the manner in which the exception has been implemented by the Member States may cause difficulties (and cross-border issues), particularly if it is limited to "extracts of works" or only benefits certain kinds of beneficiaries. This illustrates a lack of harmonization and raises the question of its mandatory character;

- regarding mentioning of the source, this second condition of Article 5(3)(a) – as it stands – does contain significant flexibility (safeguard clause in case of impossibility) to allow data analysis. Although it constitutes a burden, it cannot be considered as such as a barrier to data analysis. One may however wonder if authors actually benefit from this burden;

- it has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion.

(ii) Use of the "structure of the database" for scientific research - Article 6.2.b) of the Database Directive

1. The principles

In the following paragraphs, we will assess whether data analysis activities undertaken on databases could be covered by the exception to copyright for scientific research contained in Article 6.2.b) of the Database Directive which applies not to the works contained in a database but to the database itself (more exactly, to its structure, or to its selection and arrangement) and which provides:

> *"2. Member States shall have the option of providing for limitations on the rights set out in Article 5 in the following cases: […]*
>
> *(b) 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".*

From our previous conclusions in Part III, we know that when the data analysis is made on the basis of data and information held in a database, it will only in some cases but not often involve a reproduction or an adaptation of the selection or the arrangement of the database itself and it will almost never involve a communication (or making available) of the database to the public. For that purpose, we made a distinction between the data mining process and the data mining output(s):

- during the data mining ***process***, it may be that the selection or the arrangement of the database are copied; this will for instance be the case if all data included in the database are copied: in that case, the selection is being copied; in other cases, parts of the arrangement could also happen to be copied during the process; this will however not necessarily be the case and will depend on the circumstances and the techniques being used. Moreover, during the data analysis ***process***, there is no communication "to the public" if a database is being analysed (mined) by a group of researchers or a private company or any other "miner". A communication "to the public" does require that the work (here, the database) be communicated to a sufficiently significant number of persons;

- concerning the data mining ***output(s)*****,** we consider it most unlikely that the data analysis output will contain whole or part of a protected database which can be recognizable. This is because in the data analysis output the differences with the original database are so important that the global impression is

not the same. Moreover, the data analysis ***output*** is very unlikely, in our view, to involve a communication to the public of the copyright protected elements of the database itself (i.e. the original selection or arrangement). Considering that the data analysis output (which will in many cases be a new text, a graphic, a chart, a design, a matrix, etc.) is a new knowledge or insight, the initial data will themselves arguably not be communicated to the public via this channel, in the same way as the knowledge gained by a researcher while preparing his thesis will in itself not be found as such in the thesis (at least not in a manner which infringes intellectual property rights of other rightholders). It is indeed our view that when the output of the analysis is being communicated, such output will be composed of statistics, new relationships, new patterns, which by definition were not visible in the pre-existing works. Only really tiny elements might, by mistake or coincidence, escape the attention of the drafters of the output and appear as such in the presentation of the output or only incidentally appear as illustration.

It is not explicitly stated in the Database Directive that the exception only applies to published databases. Article 6 being a copyright exception, one may gather that it only applies to works (here, databases) which have been made public by or with the consent of the rightholder.

Divergences in implementation. The exception to **copyright** for scientific research in relation to databases contained in Article 6(2)(b) of the Database Directive has been implemented in four Member States among those considered in this Study[243]: Belgium[244], Spain[245], the UK[246] and Italy[247]. We have noticed that the requirements of "indication of the source" and "non-commercial purpose to be achieved" have been taken up in all these Member States and that the exception always applies to the "database" as a whole (and not just a part of it). Some Member States have sometimes added conditions for the exercise of this exception: In Italy, the user can "access and visualize" the database, without the author's consent, but "permanent reproduction […] shall always be subject to the rightholder's authorization"[248]. In Belgium, "the author's name and the title of the database must be mentioned", while the requirement that the use must be for the "sole" purpose of scientific research is missing. Moreover, the Belgian law draws a distinction between the reproduction on paper or any similar medium and the reproduction in any medium other than paper or similar medium, the latter aiming clearly at the reproductions in a digital format[249]. Furthermore, the implementation of the exception generally follows the Directive and combines in one provision the objectives of teaching and research without treating them differently. Except in the UK, where the exception for research is not included in the education exception.[250]

Other Member States – the Netherlands[251], Germany[252], Poland[253], Luxembourg[254], Denmark[255] and Hungary[256] – have not implemented the exception for scientific research to the copyright protection of databases contained in Article 6(2)(b) of the Database Directive. These Member States make no explicit

[243] Germany, France, the UK, Italy, Spain, Poland, Denmark, Hungary and the Benelux.

[244] Article 22bis of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[245] Article 34 of the Law No. 5/1998 of March 6, 1998 on the Incorporation in Spanish Law of Council Directive 96/9/EEC of March 11, 1996 on the Legal Protection of Databases.

[246] Article 8 of the Copyright and Rights in Databases Regulations of 18 December 1997.

[247] Article 64 sexies of the Italian Legislative Decree of 6 May 1999 No. 169 relating to the implementation of Directive 96/9/EC on the legal protection of databases.

[248] Article 64 sexies of the Italian Legislative Decree of 6 May 1999 No. 169 relating to the implementation of Directive 96/9/EC on the legal protection of databases.

[249] Article 22bis of the Belgian Law of 30 June 1994 related to Copyright and Neighboring Rights.

[250] Article 8 of the UK Copyright and Rights in Databases Regulations of 18 December 1997.

[251] The Dutch Copyright Act of 1912.

[252] The German Act of 9 September 1965 on Copyright and Related Rights.

[253] The Polish Law No. 83 of 4 February 1994 on Copyright and Neighboring Rights.

[254] The Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[255] The Danish Consolidated Act No. 202 of 27 February 2010 on Copyright.

[256] The Hungarian Act LXXVI of 22 June 1999 on Copyright.

reference to databases in their copyright legislations. In France[257], the French Intellectual Property Code provides explicitly that: "the [general copyright protection] applies to authors […] of databases which, by choice or by virtue of the arrangement of the data, constitute intellectual creations"[258]. The general copyright regime is applicable to databases accordingly (including the exception for research).

Conditions. In the event that the data analysis process involves an act of copying of the "structure of the database", this will not infringe the authors' rights if the user can prove that (Article 6.2.b) of the Database Directive):

(i) the database is used for the sole purpose of scientific research;

(ii) the sources are indicated;

(iii) the database is used to the extent justified by the non-commercial purpose to be achieved.

The three-step test also applies to the condition (article 6.3. of the Database Directive).

2. Application to data analysis

We will try to apply these three conditions to data analysis in the following paragraphs:

a. The database is used for the sole purpose of scientific research

Scientific research. The conditions of Article 6(2) of the Database Directive are similar (although not totally identical) to the conditions of Article 5(3)(a) of the Infosoc Directive (see *supra*). **Illustration**. We have seen that there is a trend in the literature and the case-law to consider that "illustration" is linked to teaching and not to research. We assume that the same principle applies here. If one indeed accepts that the exception does not only apply "to illustrate" scientific research (in which case it would become more or less useless for data analysis), the "scientific research" condition is useful for data analysis. It does exclude "non scientific" research or data mining projects which do not even qualify as "research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research projects" but this does not mean that the distinction is not adequate.

Sole purpose. Moreover, scientific research must be the sole purpose behind the use of the database for data analysis. Data analysis would fall outside of the scope of the exception from the moment it is done also for other purposes that scientific research, i.e. statistical, behavioural analysis (which do not qualify as scientific research but have essentially a commercial purpose), etc[259]. This condition may constitute a barrier for some data analysis projects and (as for the exception in the InfoSoc Directive, one may wonder why the existence of other purposes (even if non commercial) should *ipso facto* render the exception not applicable. This raises the question as to whether the "sole purpose" condition should not be better replaced by a "main purpose" condition; we will come back to this when discussing recommendations in Part IX – yet the observation concerns the exception for scientific research in general, and not just its application or relevance for TDM.

[257] The French Intellectual Property Code of 1 July 1992.

[258] Article L. 112-3 of the French Intellectual Property Code of 1 July 1992.

[259] It undeniable that the commercial/non-commercial distinction could also help in eliminating certain projects from the benefit of the exception, yet the "scientific research" condition" and the "non-commercial purpose" condition do not fully overlap, be it simply because there are obvious "commercial scientific research projects" (by the pharmaceutical sector, just to give one example).

b. The sources are indicated

Contrary to Article 5(3)(a) of the Infosoc Directive, Article 6(2) of the Database Directive imposes the indication of the source of the database but does not provide for a safeguard clause if "this turns out to be impossible".

M. Walter and S. Von Lewinski explain that:

> *"There are only two differences in the wording of Article 5(3)(a) of the Infosoc Directive and Article 6(2)(b) of the Database Directive, and they are unlikely to transpose into a difference in meaning. First the obligation in the Infosoc Directive to indicate not only the source but also the author's name may be interpreted as having a declaratory nature, given the obligation under the Berne Convention (by which all EC Member States are bound) to provide for authors' moral rights to claim authorship. Secondly, the exception from this obligation in the Infosoc Directive in cases in which such indication turns out to be impossible may also be interpreted as being of a declaratory nature, because it is a general principle of law that law never can oblige anyone to do anything that is impossible (impossibilium nulla est obligatio)".*[260]

We are not certain we can follow the second part of this opinion. In the InfoSoc Directive and in the Database Directive, the law "does not oblige anyone to do anything that is impossible" but provides for an exception and grants the benefit of it under certain conditions. If one of these conditions cannot be met, then the exception cannot be invoked. The fact that the Database Directive does not include a safeguard clause ("unless this turns out to be impossible") does mean a difference, in our view: if the user is not able to mention the sources, the legal consequence is that he cannot invoke the benefit of the exception.

Mentioning "the source" also implies mentioning "the author's name", even if this is not stated explicitly in the Directive, since we are here dealing with copyright protected databases (the author has a moral right of authorship).

As for the exception in the InfoSoc Directive, and for the same reasons as explained above, one might also, in our opinion, question the relevance of this obligation in the context of data analysis.

The wording of Article 6(2) poses challenges. While it can already be complicated to indicate the source of one database, data analysis can sometimes involve the processing of hundreds or thousands of databases. The difficulty to mention the source of all databases for data analysis can therefore be a heavy burden for the market actors. Consequently, in the absence of the reference to the terms "unless this turns out to be impossible" in Article 6(2), researchers must be able to indicate the sources (and list the sources in order to publish them) or refrain from copying them in their research project.

c. The database is used to the extent justified by the non-commercial purpose to be achieved

As for the corresponding exception in the InfoSoc Directive, only acts accomplished with a non-commercial purpose and justified by the non-commercial purpose will benefit from the exception. It is commonly admitted that "commercial" should be read as including direct and indirect economic and commercial advantages[261].

What we said about this criterion in the Infosoc Directive applies here as well (see *supra*).

It has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion.

[260] WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1042.

[261] WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1045.

3. Preliminary conclusions

As preliminary conclusions on the copyright exception of scientific research for databases:

From our previous conclusions in Part III, we know that when the data analysis is made on the basis of data and information held in a database, it will only in some cases but not often involve a reproduction or an adaptation of the selection or the arrangement of the database itself and it will almost never involve a communication (or making available) of the database to the public.

In the event that the data analysis process involves an act of copying of the "structure of the database", this will not infringe the author's rights if the user can prove that (Article 6.2.b of the Database Directive):

(i) the database is used for the sole purpose of scientific research;

(ii) the sources are indicated;

(iii) database is used to the extent justified by the non-commercial purpose to be achieved.

It is not explicitly stated in the Database Directive that the exception only applies to published databases. Article 6 being a copyright exception, one may gather that it only applies to works (here, databases) which have been made public by or with the consent of the rightholder.

Regarding the "**scientific research**" condition, if one accepts that the exception does not only apply "to illustrate" scientific research (in which case it would become more or less useless for data analysis), the "scientific research" condition is useful for data analysis. It does exclude "non scientific" research or data mining projects which do not even qualify as "research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research projects" but this does not mean that the distinction is not adequate.

Scientific research must be the **sole purpose** behind the use of the database for data analysis: data analysis would fall outside of the scope of the exception from the moment it is also done for other purposes than scientific research, i.e. statistical, behavioural analysis, etc. This condition may constitute a barrier for some data analysis projects and (as for the exception in the InfoSoc Directive, one may wonder why the existence of other purposes (even if non commercial) should *ipso facto* render the exception not applicable. This raises the question as to whether the "sole purpose" condition should not be better replaced by a "main purpose" condition.

Regarding the **mentioning of the source**, the wording of Article 6(2) poses challenges. While it can already be complicated to indicate the source of one database, data analysis can sometimes involve the processing of hundreds or thousands of databases. The difficulty to mention the source of all databases for data analysis can therefore be a heavy burden for the market actors. In the absence of a reference to the terms "unless this turns out to be impossible" in Article 6(2), researchers must be able to indicate the sources (and list the sources in order to publish them) or refrain from copying them in their research project. One might, in our opinion, question the relevance of this obligation in the context of data analysis.

Regarding the **non-commercial purpose**, it has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion.

c) Normal use of the "structure of the database" by the lawful user – Article 6.1 of the Database Directive

In the following paragraphs, we will assess whether data analysis activities could be covered by the exception to copyright for normal use by the lawful user contained in Article 6(1) of the Database Directive:

> *"1. The performance by the lawful user of a database or of a copy thereof of any of the acts listed in Article 5 which is necessary for the purposes of access to the contents of the databases and normal use of the contents by the lawful user shall not require the authorization of the author of the database. Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part".*

Although it was not specifically referred to in the Terms of Reference, this Article is relevant for data analysis for the reasons stated below.

Compulsory and not-waivable. It is the only compulsory exception to copyright in the Database Directive, the others are optional. Not only is it compulsory for the Member States but it is also an exception that cannot be waived by contract[262]. The exception was inspired from a corresponding provision in article 5.1. of the Software Directive : the idea behind it was that using a database (like using a software) does imply certain reproductions (be it in cache) and that it did not make sense to make such additional reproductions dependent upon an additional authorization from the rightholder if such rightholder had given access to the database and allowed use of it[263].

Lawful user. The term "lawful user" is not defined in the Database Directive. The definition of "lawful user" is controversial. E. Derclaye distinguishes three interpretations: 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":[264]

> *"The notion of lawful user is crucial as it determines whether only licensees or acquirers of databases or anyone can exercise the exceptions. Despite the concept's importance, it is defined neither in the Directive nor in the Recitals. Only the Explanatory Memorandum gives a definition of the lawful user: "a person having acquired the right to use a database".*[265]

E. Derclaye explains that:

> *"The third interpretation seems preferable in view of its adoption by many authors, the fact that it fits the wording of the Directive and its preparatory texts and is in accordance with the interpretation to be given to the terms "lawful user" in the Software Directive (Article 5 and 6). In conclusion, the lawful user always needs a contract: she is the person who acquired the database or a licit copy of it in a lawful way. This interpretation fits off-line databases in analog or electronic form as well as online databases only accessible through a subscription agreement restricting access through the means of, for example, a password. What about online databases accessible without payment or password, that is, websites with unrestricted access? It is submitted that, in those cases, there is an implied license by the database maker to use the database within the limits of the Directive. In other words, for online database, the freedom of access satisfies the condition of lawfulness. Thus users having lawful access to the internet are lawful users".*[266]

---

[262] See indeed article 15 of the Directive ("Binding nature of certain provisions"):

"Any contractual provision contrary to Articles 6 (1) and 8 shall be null and void".

[263] See A. STROWEL & J.P. TRIAILLE, op. cit., p. 274.

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

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

[266] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 125.

We share this view, except when this commentator states (above) that the lawful user "always needs a contract", be it thus even an implied license. Where no contractual conditions have been imposed and a website is freely accessible, we do not think there exists an implied license, but rather that no license is needed in the first place.

The lawful user needs a contract, except when the freedom of access to the database satisfies the condition of lawfulness.

In Part III, we identified four levels of access to data in the context of data analysis, and we qualified four types of data accordingly:

1. "**all to all**" access refers to Web data: as explained above, freedom of access to Web data contained in an online database satisfies the condition of lawfulness.
2. "**many to many**" access refers to Social networks data: the level of access will depend on the private account settings of the users. Accessing data by bypassing such settings would be by an "unlawful user"; according to Article 6(1) of the Database Directive: " (…) Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part" and unauthorized uses would thus be by an unlawful user;
3. "**one to many**" access refers to Contractual/publishers data: in that case, the user will have agreed to a contract and the contract will define what is lawful or not[267]; again, according to Article 6(1) of the Database Directive: " (…) Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part" and unauthorized uses would thus be by an unlawful user;
4. "**one to one**" access refers to Confidential data made available in the context of a non-disclosure agreement (not in the scope of this Study). In such case, there will also be a contract but also the data have not been published and thus the exception might not be applicable (yet this is controversial). In any case, according to Article 6(1) of the Database Directive: "(…) Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part", so that unauthorized uses would thus be by an unlawful user.

The majority of the users of a database protected by the *sui generis* right will be lawful users. E. Derclaye explains that:

> *"The interpretation of a lawful user as a lawful acquirer leaves us with very few unlawful users. The majority of users will be lawful: users generally have access to databases through the internet, libraries, as an employee or student through a subscription made by the employer or the university or by simply borrowing a database privately or acquiring the database new or second-hand for a price or free of charge. Unlawful users will be those stealing a database incorporated in a tangible medium, those subsequently acquiring a stolen or an infringing database (a pirate copy). It could also be the persons using someone's password to access a database or using, knowingly or not, a website pirating a protected database".*[268]

Borderline cases. There can be borderline or more difficult situations.

We have identified (and will analyze below) two such situations: (i) contractual conditions available on a website but not binding for the users due to a lack of consent of the users, and (ii) websites which do not require a password but which are protected by technical protection measures (TPM). For the purpose of our analysis, we assume that these websites are or contain databases protected by the *sui generis* right.

- **Terms of Use**: The "Terms of Use" of a website often provide for specific access conditions to the website applicable to the users. However, it can be argued that the user of a website does not

[267] Subject to the proviso that an exception could be made imperative and not overrideable by contract. See below on this discussion.

[268] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 125. We are not talking here about the exception of article 6.1. but about other exceptions provided for by the Database Directive.

give his/her consent to the "Terms of Use" of the website when these terms are simply available on the Home Page of said website without the user being obliged to accept "clickwrap terms and conditions". No contract between the user and the webmaster is concluded. In other words, a researcher doing data analysis on the website must be considered as a "lawful user", irrespective of the content of the Terms of Use of the websites (and regardless of the fact that these terms and conditions might prohibit data mining). The situation would be different if the webmaster had organized a "tick-box" or any other technical mechanism aiming at obliging the user to take knowledge of the content of the Terms of Use governing the website and to "accept" them explicitly.

- **TPM**: It can also happen that a user tries to get access to a website which does not require a password but which is protected by technical protection measures (TPM). This is the case, for example, for websites which have implemented a system that blocks repeated queries (i.e. crawling) from the same IP addresses in order to avoid repeated extraction of the content of the website. In this scenario, what would happen if some users have e.g. created many virtual IP addresses in order to circumvent the TPM and still be able to access the data and download them? In our opinion, such users are not "lawful users". Arguably, such use would also not qualify as "normal use" as per Article 6(1) of the Database Directive.

A question may be raised as to whether a user will be a lawful user if he/she can invoke an existing exception in order to copy the database. This will be true only if no agreement has been concluded which revokes (implicitly or explicitly) the applicability of that exception. Obviously, such possibility (of an agreement setting the exception aside) only exists to the extent that the exceptions are not considered as imperative (not waivable). In the absence of any precision to that effect in the Directive, one cannot come to the conclusion that the exceptions are unwaivable, and the existing case law of the ECJ on copyright exceptions does not allow to come to a different conclusion.

To sum up on the notion of "lawful user", the lawful user is a user who can invoke either a contractual authorization to use the database or a legal exception, provided that (1) in the latter case, no agreement which he has consented to prevents the application of the exception and that (2), in all cases, he did not circumvent any technical system meant to limit or prevent his use.

Normal use of the contents. Coming back to Article 6(1) of the Database Directive, one should then ask what "normal use of the contents" (by the lawful user) means in said article:

> *"1. The performance by the lawful user of a database or of a copy thereof of any of the acts listed in Article 5 which is necessary for the purposes of access to the contents of the database and* ***normal use of the contents*** *by the lawful user shall not require the authorization of the author of the database. Where the lawful user is authorized to use only part of the database, this provision shall apply only to that part".*

Paragraph 34 of the Explanatory Memorandum is useful in this regard:

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

This must clearly be understood as meaning that normal use takes into account the "purposes" and the "way of use and access" as set out in the agreement with the rightholder. In the absence of agreement, the intentions of the rightholder should, if known, probably come into play but this in our opinion can be debated[269].

[269] See our comment on the first of two borderline cases identified above. It is uncertain whether a rightholder may "impose" purposes and modalities of access of use by merely mentioning them in terms and conditions which do not have to be accepted before accessing the database.

Paragraph 34 also implies that the exception only applies to databases which "the rightholder has chosen to make available (..)", so it only applies to published databases.

Article 5.1 of the Software Directive, from which Article 6(1) of the Database Directive is inspired, provided as follows:

> *"1. In the absence of specific contractual provisions, the acts referred to in points (a) and (b) of Article 4(1) shall not require authorisation by the rightholder where they are necessary for the use of the computer program by the lawful acquirer in accordance with its intended purpose, including for error correction."*

One must combine the intended purpose[270], the lawful acquisition and the existence (or absence) of contractual provisions.

The exception cannot be waived by contract (see Article 15 of the Database Directive), but the contract may limit the purposes and modalities of access. As for the exception of Article 5.1 of the Software Directive, some commentators explained that the exception was therefore of limited utility to users[271].

"Access to the contents" is also mentioned in Article 6(1) ("for the purposes of access to the contents of the database") but it must be combined with "normal use" ("for the purposes of access to the contents of the database and normal use of the contents"): the act must therefore be necessary for the access *and* (not "or") the normal use. "Access" seems more neutral and less normative than "normal" use, but considering that normal use also has to be one of the purpose, the more restrictive condition '(i.e. "normal use") of the two will be decisive.

Normal use and lawful user. It is our understanding that "normal use" and "lawful user", both used in the Article 6(1) refer to the same principle and are largely redundant, i.e. the use must be in accordance with either a legal exception or a contractual license, while the legal exception cannot as such be waived by the contract but the contract may limit the purpose and modalities of access; further, the use may not amount to circumventing a technical system put in place by the author.

The acts are only authorized to the extent that they are "**necessary** for the purposes of access to the contents of the database and normal use of the contents" (our emphasis). This requirement will in our opinion exclude most data analysis operations from the benefit of this provision: it is unlikely that data analysis will be "necessary" to access the contents and to use such content in a normal manner; arguably, doing data analysis on a database has as its purpose to extract "new patterns", "new insights", which is not what would normally be the purpose in the context of a normal use of a database: data analysis intends, by "mining" the contents, to extract new information: this, we would think, does not fall within the "normal use" of a database. So, this condition does constitute a significant obstacle to data analysis. The only example one could think of are databases which have been put on the market precisely for data mining purposes, i.e. where the normal use of the database will be to mine the contents. In such case, data analysis is necessary, amounts to a normal use and the user is a lawful user. And requiring an additional authorization from the rightholder for such data mining would, in such case, be "absurd"[272].

Last condition, the act (reproduction) will only be lawful if it is done for the purposes of access (…) and normal use *by the lawful user*. The access and the use must thus arguably be made by the user itself, not by a third party authorized by the lawful user.

[270] In the French version of the Directive : « d'une manière conforme à sa destination »

[271] See A. STROWEL & J.P. TRIAILLE, op. cit., p. 274.

[272] See A. STROWEL & J.P. TRIAILLE, op. cit., p. 274.

Conclusion on the "normal use" exception for databases. Where does all this leave us when applied to data mining activities? It does not give much room for data mining. It will only be authorized under this provision if it is "*necessary*" to have access to the contents and for its normal use; if the activity requires bypassing some technical protection systems, it cannot benefit from this exception. It also means that if the license or subscription agreement prohibits data mining, such prohibition is binding on the user, provided he has accepted the condition (and his consent thereto can be proven).

d) Conclusions on the exceptions to copyright

See below in this Study: we have assembled together, under Part V, A, our conclusions on the copyright exceptions and on the *sui generis* exceptions.

B. EXCEPTIONS TO THE *SUI GENERIS* RIGHT IN THE DATABASE DIRECTIVE

After having examined various exceptions to copyright both in the InfoSoc Directive and in the Database Directive and assessed whether they are useful for data mining purposes, we will now examine the existing exceptions to the *sui generis* right of the maker of a database, as such right has been introduced in EU legislation by the Database Directive.

We will successively deal with the exception for the lawful user – under a), and the extraction of data for scientific research – under b).

a) Extraction of "insubstantial parts" by the lawful user - Article 8.1 of the Database Directive

Article 8.1 defines the rights and obligations of the lawful user of a database protected by the *sui generis* right. Although it was not specifically referred to in the Terms of Reference, this Article is in our opinion worth examining for data analysis for the reasons stated below. In the following paragraphs, we will assess whether data analysis activities could fall within the rights recognized to the "lawful user" by Article 8.1 of the Database Directive:

> *"The maker of a database which is made available to the public in whatever manner may not prevent a lawful user of the database from extracting and/or re-utilizing insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever. Where the lawful user is authorized to extract and/or re-utilize only part of the database, this paragraph shall apply only to that part".*

The expression "**lawful user**" has been described above; it must be understood in the same manner here.

One should note in passing that this provision is somewhat illogical: it reserves to the lawful user the benefit of extracting insubstantial parts, while the right of extraction only applies, as it is defined, to all or substantial parts of the contents of the database. By consequence, since extracting insubstantial parts do not fall within the scope of the extraction right, the maker of the database has no exclusive right on acts of extractions of insubstantial parts. This possibility should thus be open to both lawful users and unlawful users – subject only to the application of legislation on authorised access to computer systems or hacking. But this is another matter which does not fall directly within the scope of this Study.

The article is different from Article 6.1 of the Database directive which lays down a similar exception to copyright: the former provision refers to the lawful user but authorizes him to extract and reutilize "**for**

any purpose whatsoever": this is thus broader than the notion of a "normal use" under Article 6.1. "Any purpose whatsoever" includes purposes which the maker may not have intended, including "non normal" uses.

The database must have been **made available to the public "in whatever manner**": this, in our view, refers to the idea that the provision does not apply if the database has not been made public (and is thus, for example, a purely internal database or a confidential one)[273]. If the database is not freely accessible but is only accessible on subscription, then users will only be "lawful" if they can avail themselves of the authorization to access the database granted through the subscription and remain within the limits of use authorized by the agreement.

The terms "insubstantial part" and "quantitative" and "qualitative" are not defined in the Database Directive. The Database Directive defines "extraction" in Article 7(2)(a) 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". Therefore, the extraction of insubstantial part of a database, provided by Article 8.1, does not fall under the definition of "Extraction" in Article 7(2)(a).

In the BHB case[274], the ECJ has given the following definitions to these terms:

> *§69.[…] It appears from that recital that the assessment, in qualitative terms, of whether the part at issue is substantial, must, like the assessment in quantitative terms, refer to the investment in the creation of the database and the prejudice caused to that investment by the act of extracting or re-utilising that part.*
>
> *§70. The expression 'substantial part, evaluated quantitatively', of the contents of a database within the meaning of Article 7(1) of the directive refers to the volume of data extracted from the database and/or re-utilized, and must be assessed in relation to the volume of the contents of the whole of that database. If a user extracts and/or re-utilises a quantitatively significant part of the contents of a database whose creation required the deployment of substantial resources, the investment in the extracted or re-utilised part is, proportionately, equally substantial.*
>
> *§71. The expression 'substantial part, evaluated qualitatively', of the contents of a database refers to the scale of the investment in the obtaining, verification or presentation of the contents of the subject of the act of extraction and/or reutilisation, regardless of whether that subject represents a quantitatively substantial part of the general contents of the protected database. A quantitatively negligible part of the contents of a database may in fact represent, in terms of obtaining, verification or presentation, significant human, technical or financial investment.*
>
> *[…] §73. It must be held that any part which does not fulfil the definition of a substantial part, evaluated both quantitatively and qualitatively, falls within the definition of an insubstantial part of the contents of a database.*

This interpretation has been commented by E. Derclaye who explains that:

> *"With its interpretation of the terms "substantial part" and "insubstantial part", the Court clearly makes a link between the substantial investment and the infringement test. There will only be infringement if the substantial investment is harmed by the act of the user. Thus, if the user takes a substantial part which does not represent the substantial investment of the database maker, the*

[273] This requirement is not mentioned in the corresponding Article 6.1., but since the latter deals with copyright protected databases, one may suppose that exceptions only play vis-à-vis databases which the author has decided to disclose (moral right of divulgation). On the *sui generis* right, see Paragraph 34 of the Explanatory Memorandum which is useful to this regard:

> *"(34) Whereas, nevertheless, once the rightholder has chosen to make available a copy of the database to a user,* ***whether by an on-line service or by other means of distribution****, that lawful user must be able to access and use the database for the purposes and in the way set out in the agreement with the rightholder, even if such access and use necessitate performance of otherwise restricted acts;" (*our emphasis*).*

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

> *investment is not harmed and there cannot be infringement, even if there was an act of extraction or reutilization. This link has been suggested by a number of commentators and transpires from recital 42 of the Database Directive[275], to which the Court refers."*[276]

Consequently, data analysis will fall within the rights recognized to the lawful user by Article 8.1 and will not infringe the rights of the database maker if the researcher ("lawful user") extracts insubstantial parts of the database. This must be assessed on a case-by-case basis.

The rights conferred by Article 8(1) are however not absolute. Article 7(5) provides an exception to the general rule contained in Article 8(1) and states that:

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

This provision prohibits the repeated and systematic extraction of insubstantial parts of the content of the database. This can be relevant for data analysis. In the BHB case[277], the ECJ explains that the two conditions "repeated" and "systematic" apply cumulatively:

> *"§86. The purpose of Article 7(5) of the directive is to prevent circumvention of the prohibition in Article 7(1) of the directive. Its objective is to prevent repeated and systematic extractions and/or re-utilisations of insubstantial parts of the contents of a database, the cumulative effect of which would be to seriously prejudice the investment made by the maker of the database just as the extractions and/or re-utilisations referred to in Article 7(1) of the directive would".*

In the BHB case[278], the ECJ also defines what is meant by "acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database" (in the sense of article 8.1.) and states that:

> *"§89. Under those circumstances, 'acts which conflict with a normal exploitation of [a] database or which unreasonably prejudice the legitimate interests of the maker of the database' refer to unauthorised actions for the purpose of* ***reconstituting****, through the cumulative effect of acts of extraction,* ***the whole or a substantial part of the contents*** *of a database protected by the sui generis right and/or of making available to the public, through the cumulative effect of acts of re-utilisation, the whole or a substantial part of the contents of such a database, which thus seriously prejudice the investment made by the maker of the database" (our emphasis).*

The interpretation given to Article 7.5 by the ECJ is highly relevant for data analysis. E. Derclaye explains that:

> *"The ECJ's interpretation of Article 7.5 is simple: the extractions or reutilizations must have the effect of* ***reconstituting*** *a substantial part or the entire database. […] "Repeated and systematic" simply means that the user's acts* ***reconstitute*** *the whole or a substantial part of the maker's database. By implication, because the accumulation of insubstantial parts must amount to a substantial part, the extraction or reutilization conflicts with a normal exploitation of the database or unreasonably prejudices the legitimate interests of the database maker. […] Like Article 7.1, Article*

[275] Recital 42 provides that "Whereas the special right to prevent unauthorized extraction and/or re-utilization relates to acts by the user which go beyond his legitimate rights and thereby harm the investment; whereas the right to prohibit extraction and/or re-utilization of all or a substantial part of the contents relates not only to the manufacture of a parasitical competing product but also to any user who, through his acts, causes significant detriment, evaluated qualitatively or quantitatively, to the *investment* (emphasis added)".

[276] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 111.

[277] ECJ, 9 November 2004, case C-203/02, *The British Horseracing Board Ltd and Others v William Hill Organization Ltd., par.* 85 et s.

[278] ECJ, 9 November 2004, case C-203/02, The British Horseracing Board Ltd and Others v William Hill Organization Ltd., par. 89.

> *7.5 is an objective test: it is whether the user in effect takes a substantial part which incorporates a substantial investment. The harm caused cannot be potential; it must exist." (*our emphasis*)*

While Article 8.1 opens significant possibilities for data analysis (limited however to lawful users), Article 7.5 may first appear to constitute a barrier thereto, because data analysis generally involves repeated and systematic extractions of a database.

However, beyond the wording of Article 7.5 which may appear restrictive at first sight, the ECJ's interpretation of Article 7.5 is favourable for data analysis. Indeed, the ECJ has made clear that the repeated and systematic extractions are prohibited to the extent that they aim at "**reconstituting** the whole or a substantial part of the contents of a database protected by the *sui generis* right which thus seriously prejudice the investment made by the maker of the database". Moreover, as explained by E. Derclaye, the harm caused cannot be potential, it must exist.

The purpose of data analysis is to obtain new insights from the content of the database. The fact that such use could harm the investment made by the database maker is questionable. This must be assessed on a case-by-case basis. In our opinion, the purpose of data analysis is not to "reconstitute" (as required by the ECJ) the whole or a substantial part of a database. Furthermore, the output of the process will in most cases not reproduce (nor, therefore, "reconstitute") any part or any data from the databases which have been used as sources.

As a **conclusion on the right to extract insubstantial parts**, Article 7.5 should, in most cases, not be an obstacle, and Article 8.1 will provide a significant "pass through" for "lawful users" of databases willing to proceed to data analysis. Lawful users are users who have a legitimate access to a database, either through a contract (which does not prohibit data analysis) or through an exception.

In practice, if the subscription agreement to a database does not prohibit data analysis, the user may perform such analysis if, in order to do so, he only extracts insubstantial parts (quantitatively or qualitatively). Repeated and systematic extractions are only prohibited if they aim at reconstituting the whole or a substantial part of the database.

If the agreement prohibits data analysis, the user may not invoke the right granted by article 8.1.: he would become an unlawful user.

b) Extraction of "data" for scientific research - Article 9.b of the Database Directive

1. <u>The principles</u>

In the following paragraphs, we will assess whether data analysis activities could be covered by the exception for scientific research to the *sui generis* right contained in Article 9(b) of the Database Directive:

> *"Member States may stipulate that lawful users of a database which is made available to the public in whatever manner may, without the authorization of its maker, extract or re-utilize a substantial part of its contents:*
>
> *[…] (b) in the case of extraction for the purposes of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved".*

As for Article 8.1., the database must have been **made available to the public "in whatever manner**": this, in our view, refers to the idea that the provision does not apply if the database has not been made

public (and is thus, for example, a purely internal database or a confidential one)[279]. If the database is not freely accessible but is only accessible on subscription, then users will only be "lawful" if they can avail themselves of the authorization granted through the subscription.

It must be noted that Article 9 provides for an exception to the right of extraction and not to the right of re-utilization. As explained in Part III, data analysis will, in most cases, involve extraction of all or substantial parts of the contents of the database but it will normally not amount to re-utilizing the same. Therefore, Article 9 is relevant for data analysis.

Recital 51 of the Database Directive leaves to Member States the possibility to restrict the exception to the *sui generis* right to "certain categories of teaching or scientific research institutions", without stating any criteria for such a limitation.[280] We are not aware of any comprehensive study on this issue. Therefore, we cannot make conclusions in the context of data analysis. However, it must be kept in mind that the Member States have the possibility to restrict the benefit of the exception to certain categories of scientific research institutions. This, in turn, could potentially constitute a serious barrier for data analysis done by all other public institutions and/or private companies which are not "privileged" by the national legislations of the Member States.

Divergences in implementation of Article 9(b). The exception to the **sui generis right** for scientific research contained in Article 9(b) of the Database Directive has been implemented in nine countries among those considered in this study[281]: Belgium[282], Spain[283], the UK[284], the Netherlands[285], France[286], Germany[287], Poland[288], Luxembourg[289], and Hungary[290].

We have noticed that the requirements of "indication of the source" and "non-commercial purpose to be achieved" have been taken up in all these Member States and that the exception always applies to a "substantial part of the database".

Some Member States have added new conditions to the exception contained in Article 9(b) of the Database Directive: In Hungary the extraction of data for scientific research is authorized "in a manner and to the extent consistent with such purpose". In Belgium, not only the source, but also "the name of the database maker and the title of the database must be mentioned". In France, the regime is rather restrictive: (i) some databases are excluded from the benefit of the exception ("to the exclusion of databases created for educational purposes and databases created for a digital written edition), (ii) some use are excluded ("to the exclusion of entertainment or recreational activity"), (iii) the beneficiaries are limited ("so far as the public to whom the extraction and the re-utilization are intended is mainly composed

[279] This is similar to what has been said for the exception to copyright contained in Article 5(3)(a) of the Infosoc Directive with regards to "unpublished works".

[280] Recital 51: "Whereas the Member States, where they avail themselves of the option to permit a lawful user of a database to extract a substantial part of the contents for the purposes of illustration for teaching or scientific research, may limit that permission to certain categories of teaching or scientific research institution"

[281] Germany, France, the UK, Italy, Spain, Poland, Denmark, Hungary and the Benelux.

[282] Article 7 of the Law of 31 August 1998 transposing into Belgian law the European Directive of 11 March 1996 on the legal protection of databases (MB 14.11.1998).

[283] Article 135 of the Law No. 5/1998 of March 6, 1998 on the Incorporation in Spanish Law of Council Directive 96/9/EEC of March 11, 1996 on the Legal Protection of Databases.

[284] Article 20 of the UK Copyright and Rights in Databases Regulations of 18 December 1997.

[285] Article 5 of the Dutch Database Act of 8 July 1999.

[286] Article L. 112-3 of the French Intellectual Property Code of 1 July 1992.

[287] Article 87C of the German Act of 9 September 1965 on Copyright and Related Rights.

[288] Article 8 of the Polish Law on the Protection of Database of 27 July 2001.

[289] Article 68 of the Luxemburg Law of 18 April 2001 on Copyrights, Neighboring Rights and Databases.

[290] Article 84C of the Hungarian Act LXXVI of 22 June 1999 on Copyright.

of pupils, students, teachers or researchers directly involved"), and the user must pay a price ("the use of the extraction […] is compensated by a remuneration negotiated on a lump sum basis").

Furthermore, some countries do not specify that the research must be "scientific" (France and the UK) and in Germany ("personal scientific use") and in Poland ("didactic or research purposes") the exception for research is not included in the education exception. Finally, the exception for scientific research contained in Article 9(b) of the Database Directive has not been implemented in Italy and in Denmark so far.

The extraction of data, for the purpose of data analysis, will not infringe the database maker's rights when the user is the "lawful user" of the database (see *supra*), and can prove that (cumulative conditions):

(i) data are extracted for the purpose of illustration for teaching or scientific research;

(ii) the source is indicated;

(iii) data are extracted to the extent justified by the non-commercial purpose to be achieved.

2. Application to data analysis

In addition to what has already been said about the "lawful user" condition, we will analyze the fulfillment of the three conditions in the following paragraphs:

a. The data are used for the purpose of illustration for teaching or scientific research

Contrary to the exceptions to copyright contained in the Article 5(3)(a) of the Infosoc Directive and in Article 6(2) of the Database Directive, Article 9(b) of the Database Directive does not include the adjective "sole" purpose of illustration for scientific research. As a result, data analysis done partially for scientific research and for non-commercial[291] purposes other than scientific research i.e. statistical, behavioural analysis (which do not qualify as scientific research but have essentially a commercial purpose) etc. would still be covered by the exception.

As we already explained, we do not think that the requirements "scientific research" and "non commercial purposes" are redundant, for various reasons (see our explanation on this regarding the exceptions for scientific research in the InfoSoc Directive (re: Article 5.3.a) – The principles). Scientific research and non-commercial purposes are two separate conditions.

We do not think that inserting definitions, either for "scientific research", or for "commercial/non-commercial" purposes" in a legal provision would be an ideal solution: a case-by-case approach by the courts, on the basis of explanatory memorandums and preparatory works of the texts, with the assistance of legal commentators, should provide for the necessary flexibility in this regard.

The definitions of "illustration" and "scientific research" present the same uncertainties as for the exceptions contained in the Article 5(3)(a) of the Infosoc Directive and in Article 6(2) of the Database Directive. E. Derclaye explains:

> *"Another question is whether the term "illustration" relates only to teaching or also to scientific research. The wording of the article is unclear. It can mean that, in the domain of scientific research, the exception only pertains to acts of illustration. If "illustration" relates to both teaching and scientific research the exception is narrower than if it relates only to teaching".*[292]

[291] The non-commercial purpose remains a condition (see third condition).

[292] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 132.

As explained *supra*, "scientific" research excludes (we think) e.g. purely repetitive research. With regard to copyright exceptions, we have seen that there is a trend in the literature and the case-law to consider that "illustration" is linked to teaching and not to research. We assume that the same principle applies here. If this was not the case, the exception would be of almost no use to data mining.

In accordance with Recital 36 of the Directive, "the term 'scientific research` within the meaning of this Directive covers both the natural sciences and the human sciences", but the term is not further defined.

In **conclusion**, data analysis done for the purpose of scientific research, or for the purpose of scientific research and other non-commercial purposes, can benefit from this exception, which is thus very useful and relevant, provided it is made by lawful users and does qualify indeed as "scientific research".

b. <u>The source is indicated</u>

Article 9(b) imposes the indication of the source of the database. Similar to Article 6(2) of the Database Directive, and contrary to Article 5(3)(a) of the Infosoc Directive, Article 9(b) of the Database Directive does not drop this condition if "this turns out to be impossible".

E. Derclaye notes that:

> *"The Directive does not explain what indication of the source means. Probably indentifying the producer is sufficient (by analogy with art. 5.3 of the Infosoc Directive).*[293]

It is not certain to us that the identity of the maker must be mentioned: the "source" could also be the URL address or the name of the database. The identification of the source can be problematic. As explained, data analysis can sometimes involve the processing of hundreds or thousands of sources which can qualify as databases. The difficulty to mention the source of all the databases for data analysis can be a burden for the market actors (in terms of time and investment). The absence of reference to "unless this turns out to be impossible" leaves however no room for interpretation. The lawful user must indicate the source of the database.

As for the obligation under copyright to mention the name of the author, **one might also question the relevance of this obligation** in the context of data analysis, for the following same reasons:

- this requirement seems logical for the exception regarding illustration for teaching because in that case, a work is being reproduced (or communicated to the public) and is thus "visible" in the teaching materials (so that every copy of such materials involve a copy of the pre-existing works): it is thus understandable that, in such cases (like for the quotation exception) the source, including the author's name, must be indicated;
- in many scientific research projects however (and particularly in TDM projects), the research project or process may necessitate to copy pre-existing works but it may well be (and, in TDM projects, it will almost always be the case) that the research output does not include any part of the pre-existing works but only draws conclusions (new insights, new patterns) thereon (see Part III of this Study and the difference we make between the data analysis process and the data analysis output). In such situations, a copy of the output (as opposed to a copy of the teaching materials) will not involve any copy of the preexisting works. One may question whether it is justified to impose that the makers' names of all preexisting databases used for (i.e. copied during the process of) the research be mentioned[294]. Furthermore, in TDM cases, the list of "sources" risk being very long: what will be the particular interests and benefits for the maker of being mentioned? And one may wonder what would be the damage of not being mentioned amongst hundreds of others.

[293] DERCLAYE, E., "The Legal Protection of Databases: A Comparative Analysis", Edward Elgar, 2008, p. 132.

[294] This is all the more true that, for ethical or reputational considerations, the researchers will usually spontaneously cite their sources – be it to avoid being criticised by their peers.

c. The data are used to the extent justified by the non-commercial purpose to be achieved

As for the corresponding exception in the InfoSoc Directive, and as already explained above, only acts accomplished with a non-commercial purpose and justified by the non-commercial purpose will benefit from the exception. It is commonly admitted that "commercial" should be read as including direct and indirect economic and commercial advantages[295].

What we said about this criterion in the Infosoc Directive applies here as well (see *supra*).

It has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion.

3. Preliminary conclusions

As preliminary conclusions on this exception:

From our previous conclusions in Part III, we know that when the data analysis is made on the basis of data and information held in a database, from a *sui generis* point of view:

- data analysis will, in most cases, involve extraction of all or substantial parts of the contents of the database;
- but it will normally not amount to re-utilizing the same.

In the event that the data analysis process involves an act of extraction of substantial parts of contents, this will be covered by the exception if the user can prove that:

(i) he is a lawful user;
(ii) the database is used for the purpose of scientific research;
(iii) the sources are indicated;
(iv) the database is used to the extent justified by the non-commercial purpose to be achieved.

The exception only applies to databases which have been **made available to the public**.

Regarding the "**lawful user**" exception, we refer to what has been explained above.

Regarding the "**scientific research**" condition, if one accepts that the exception does not only apply "to illustrate" scientific research (in which case it would become more or less useless for data analysis), the "scientific research" condition is useful for data analysis. It does exclude "non scientific" research or data mining projects which do not even qualify as "research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research projects" but this does not mean that the distinction is not adequate.

Under Article 9b) of the Database Directive, scientific research **does not have to be the sole purpose** behind the use of the database for data analysis: data analysis remains within the scope of the exception even it is also done for other purposes that scientific research, i.e. statistical, behavourial analysis, etc. The absence of the term "sole" purpose of scientific research opens up the exception to data analysis done for scientific research but also for other objectives on the side ("*scientific research purpose plus other purposes*").

Regarding the **mentioning of the source**, the wording poses challenges. While it can already be complicated to indicate the source of one database, data analysis can sometimes involve the processing of hundreds or thousands of databases. The difficulty to mention the source of all databases for data

[295] WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1045.

analysis can therefore be a heavy burden for the market actors. In the absence of a reference to the terms “unless this turns out to be impossible”, researchers must always indicate the sources (and list the sources in order to publish them) or refrain from extracting data in their research project. One might, in our opinion, question the relevance of this obligation in the context of data analysis. The recent case-law of the ECJ might reinforce this analysis, in the sense that exceptions should be interpreted in a manner which is in conformity with their objective. What would be the objective of having an author’s name or a source be listed, in the internal servers of the data miners, amongst thousands of others, particularly where the TDM output (the report) does not quote any excerpt of any of the pre-existing works?

Regarding the **non-commercial purpose**, it has been stated on various occasions that the commercial v. non-commercial criterion is hard to apply. This may be true and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion.

c) Conclusions on the exceptions to the *sui generis* right

See below in this Study: we have assembled together, under Part V, A (hereafter) our conclusions on the copyright exceptions and on the *sui generis* exceptions.

V. IMPACT OF THE CURRENT COPYRIGHT AND DATABASE LEGAL FRAMEWORK ON DATA ANALYSIS

A. DATA ANALYSIS, COPYRIGHT AND DATABASE PROTECTION RULES

In this Part, we will summarize the findings of the previous parts of this Study. How does copyright and the *sui generis* right relate to data analysis, and in particular: (1) to what extent does data analysis fall under these exclusive rights and (2), to what extent are exceptions applicable.

Exclusive rights. We have examined in Part III to what extent the exclusive rights in both copyright (i.e. mainly the reproduction right) and in the *sui generis* right (i.e. the extraction right) constituted obstacles to data analysis and have concluded as follows:

- from a copyright standpoint:
  - when the data analysis is made on the basis of data and information which are protected by **copyright**, data analysis does, **in most cases**, involve a **reproduction** of protected materials and arguably also **translating** or **adapting** the same (with such translation or adaptation falling, in our view, within the scope of the reproduction right under the InfoSoc Directive), but not communicating them to the public nor making them available;
  - when the data analysis is made on the basis of data and information held in a database, it will only **in some cases but not often** involve a **reproduction or an adaptation** of the (structure of the) database itself and it will almost never involve a communication (or making available) of the database to the public;
- from a database protection law (*sui generis* rights) standpoint, when the data analysis is made on the basis of data and information held in a database :
  - data analysis will, **in most cases**, involve **extraction** of all or substantial parts of the contents of the database;
  - but it will **normally not amount to re-utilization** the same.

Exceptions. We then analysed in Part IV which exceptions in the InfoSoc Directive and in the Database Directive could be invoked for data analysis activities.

We concluded that:

- in the area of copyright:
  - the **temporary reproduction** exception would **only in exceptional circumstances** be a possible argument to exempt data analysis;
  - **use of works for scientific research** – with the word "scientific" excluding non-scientific research - **can legitimate data analysis** on published works, if it is its sole purpose (if other purposes, such as marketing etc., are also present, the exception may not be invoked); mentioning the source and the authors' name of the works can be a heavy burden (even if one may wonder whether this is of any utility); works can only be used ("mined") to the extent justified by a **non-commercial purpose**, but this criterion should not be interpreted too strictly (no individual assessment "work by work"); borderline cases exist between "scientific"

and "non-scientific" and between "commercial" and "non-commercial", which is not a reason to abandon these criteria;

- **use of the structure of a database for scientific research**[296] – with the word "scientific" excluding non-scientific research - **can legitimate data analysis** on the structure of published databases, if scientific research is its sole purpose (if other purposes, such as marketing, humanitarian aid, etc., are also present, the exception may not be invoked); mentioning the source is compulsory and can be a barrier (even if one may wonder whether this is of any utility); the structure of the database can only be used ("mined") to the extent justified by a **non-commercial purpose**,; borderline cases exist between "scientific" and "non-scientific" and between "commercial" and "non-commercial", which is not a reason to abandon these criteria;
- the **normal use of the structure of the database by the lawful user does not provide for interesting possibilities**, as it only applies if the acts are "necessary for the purposes of access and normal use of the contents of the database"; data analysis does not correspond to a "normal use" of a database (in the sense of the Database Directive);

- as regards the *sui generis* right:

  - the **right of the lawful user to extract insubstantial parts** of a database can be exercised for any purpose whatsoever (scientific or not, commercial or not); however, it can only be invoked by lawful users, i.e. by users who can invoke an existing exception (in which case, this right does not add anything to the existing situation) or an authorisation from the rightholder; so, **if the subscription agreement prohibits data analysis (directly or indirectly), this right is of no use;**
  - **extraction of data for scientific research** – with the word "scientific" excluding non-scientific research - **can legitimate data analysis** on published databases, even if scientific research is not its sole purpose; the exception can however **only be invoked by lawful users**; mentioning the source and the authors' name of the works can be a heavy burden (even if one may wonder whether this is of any utility); works can only be used ("mined") to the extent justified by a **non-commercial purpose**,; borderline cases exist between "scientific" and "non-scientific" and between "commercial" and "non-commercial", which is not a reason to abandon these criteria.

It is worth remembering that, apart from the temporary reproduction exception to copyright, all other exceptions mentioned above are optional. Differences in implementation (or absence of implementation in some Member States) will normally create difficulties for cross-border data analysis projects.

B. IMPACT OF THE CURRENT COPYRIGHT AND DATABASE PROTECTION RULES ON THE DIFFERENT STAKEHOLDERS

On the basis of what we analyzed up to now, one can make the following general observations:

When data analysis **does not involve works protected** by copyright or falling within the *sui generis* right, no exception need to be invoked and data analysis can be made freely – subject in certain Member States (and this is a controversial issue) to the application of parasitic behavior (amounting to unfair trade practices)[297].

When data analysis **does involve works protected** by copyright, since data analysis necessitates reproductions in the vast majority of cases:

[296] Such use will not happen often in data analysis activities (see paragraph above).

[297] See our explanations on this in Part VI of this Study.

- the rightholders are well placed to oppose it by invoking copyright;
- data miners depend upon either **their authorization or an exception**.

Copyright owners (and publishers) are thus in a comfortable position since they can refuse to grant their authorization or impose their conditions.

Users of copyright protected content seeking to engage in data analysis (research and research institutions and data analysis companies in particular) have the choice of negotiating and concluding licenses, abandoning their projects if they do not succeed in concluding a license, or engaging in mining activities without authorisation, thus taking the risk of possibly being sued (even though we found no published case-law where this happened). This lack of case-law might come from the fact that proving the use of a work or of a database in a data analysis project might not be easy : as we indicated (in Part III), in almost all cases, while the process of data mining requires copying of works or databases, the data mining output (the resulting report) will not contain any track or excerpts from the works or databases which have been used; in such cases, no part of the materials used remains visible – so that publishers might not even be able to notice that their works have been copied during the process[298].

The only other possibility for perspective miners is to try to rely on existing exceptions. This will mainly be possible for **researchers and research institutions**. This raises however the following difficulties:

- the exceptions (except the temporary reproduction one, which is of hardly no use in this context) are not harmonized throughout the EU;

- the conditions of the exceptions are submitted to various requirements which are not clear;

- companies seeking to engage in data mining for reasons other than scientific research cannot rely on exceptions which limit the possibility to do data analysis for scientific research purposes only. Data analysis for marketing purposes (predictive analysis of consumer demands, for instance) do not fall within the research exceptions, except possibly if done by research institutions and not for the purpose of proposing new products or services on the market[299]); we will consider in our recommendations (see Part IX in this Study) whether a TDM exception should also cover other research than scientific research;

- in some cases, notably under the research exception in the Infosoc Directive, scientific research must even be the sole purpose of the project: if other purposes are being pursued at the same time in addition to scientific research, if the text is to be strictly interpreted, the exception does not apply;

- mentioning the sources (sometimes, including the author's names) can be a heavy burden; in some cases, this requirement may be neglected if such mentioning proves impossible; in other cases, if it proves impossible, the exception cannot be invoked and the project must be abandoned;

- in almost all cases, the purpose of the mining may only be non-commercial in order for this activity to benefit from an exception: any commercial purpose renders the project illegal without the rightholders' consent; it seems to us that enlarging the exception to also cover commercial purposes is not desirable and would possibly be contrary to the three-step test;

- regarding the exception allowing extraction of unsubstantial parts of the contents of a database for the benefit of the lawful user (Article 8.1. of the Database Directive), whereby

[298] Except if they had recourse to raid visits on the premises of the research institution or company.

[299] The relevant distinction regarding marketing will mostly be whether it is for commercial purposes or not (with the answer, for marketing, being affirmative).

> the lawful user can do data mining both for non scientific research purposes and for commercial purposes: in this context, he may use insubstantial parts of the contents of databases; he may also do repeated and systematic extractions of insubstantial parts – which would normally amount to extracting a substantial part - provided that the purpose is not to reconstitute the whole or a substantial part of the database. This may seem like an exception which on its face could cover a number of data analysis activities, but as it is (quite illogically)[300] limited to lawful users, it means that the subscription agreement may prohibit data mining (either explicitly, or implicitly by prohibiting acts which are necessary to achieve data mining); the exception therefore remains totally dependent upon the existence of a previous contractual agreement between rightholders/publishers and users/miners. Also, it remains to be seen if any significant TDM project can be achieved by simply relying on insubstantial parts of databases.

The exceptions for the benefit of users who would like to engage in data analysis are all **waivable exceptions**: subscription agreements to databases of works can limit or prohibit data analysis, either explicitly or indirectly (by prohibiting certain acts which are in practice necessary to perform data analysis).

All in all, our evaluation at this stage of our analysis is that rightholders and publishers are in a more comfortable position than research institutions, let alone commercial companies willing to engage in data analysis projects. Whether this is a desirable result or not is a policy option, for which the objectives of the European Union to promote research and innovation should be taken into account.

[300] This seems illogical to limit such rights on insubstantial parts because the definition of extraction refers to substantial parts, so that extraction insubstantial parts should fall outside of the extraction part altogether and thus be open to all users, “lawful or not”.

VI. OTHER LEGAL PROVISIONS RELEVANT TO DATA ANALYSIS

We will assess whether there are other legal provisions other than those dealing with exclusive rights or with exceptions under copyright and database protection rules which are relevant to data analysis in the European Union[301].

We did not identify, as asked in the Terms of Reference, other general provisions, on "research purposes", for example: such provisions on research are incorporated in the copyright or database protection legislation (and were already dealt with in this Study).

We could maybe add one provision from within copyright, i.e. the exception for the purposes of administrative proceedings, judicial proceedings, etc., which we thought should be mentioned for the sake of completeness but which does not represent a major importance for the private sector (the exception mainly benefit the police, justice and investigation services) and corresponds to a very specific sector[302].

The purpose of this section is to describe legislation, from various fields of law, which may have an impact on data analysis projects. This is not an exhaustive inventory and by listing them we do not suggest that they should be adapted. The idea is to try and give a general overview of the rules which may impact data analysis.

One could find inspiration in e.g., the InfoSoc Directive and the Database Directive which both provide that they apply without prejudice to certain legal provisions. Those legal provisions are indicative of certain other legislation or legal principles which can be seen sometimes as facilitators sometimes as impediments to text and data mining. We reproduce the two (very similar) articles (both entitled "Continued application of other legal provisions"):

- The InfoSoc Directive mentions in particular (art. 9) : "*patent rights, trade marks, design rights, utility models, topographies of semi-conductor products, typefaces, conditional access, access to cable of broadcasting services, protection of national treasures, legal deposit requirements, laws on restrictive practices and unfair competition, trade secrets, security, confidentiality, data protection and privacy, access to public documents, the law of contract*";

- Article 13 of the Database Directive refers to: "*copyright, rights related to copyright or any other rights or obligations subsisting in the data, works or other materials incorporated into a database, patent rights, trade marks, design rights, the protection of national treasures, laws on restrictive practices and unfair competition, trade secrets, security, confidentiality, data protection and privacy, access to public documents, and the law of contract.*"

[301] The authors wish to thank Prof. Cécile de TERWANGNE, from the University of Namur, for her useful remarks on these issues. Prof. de TERWANGNE wrote her PhD on the reuse of public sector information and is one of the European experts in the field.

[302] In the InfoSoc Directive, see Article 5.3.e:

> *"(e) use for the purposes of public security or to ensure the proper performance or reporting of administrative, parliamentary or judicial proceedings".*

In the Database Directive, see Article 6.2.c regarding copyright:

> *"(c) where there is use for the purposes of public security of for the purposes of an administrative or judicial procedure".*

And in the same Directive, see Article 9.c regardind the sui generis right:

> *"(c) in the case of extraction and/or re-utilization for the purposes of public security or an administrative or judicial procedure".*

Some of these legal provisions are relevant for this Study, some are not; others were not mentioned in these articles but will be mentioned hereunder. We tried to list the most relevant legislation which may constitute obstacles to data analysis.

As was just stated, it is important to note that, by mentioning them, we are not thereby suggesting that these legislations should be adapted or that they constitute obstacles to TDM.

A. LEGAL PROVISIONS AND RECENT DEVELOPMENTS ALREADY MENTIONED (*REMINDER*)

We already mentioned above legal provisions or recent voluntary initiatives which have as result that they facilitate access to data and thereby also data analysis: Open Access initiatives, Creative Commons licenses and the Directives on the re-use of public sector information (PSI). We may refer to the explanations given on this in Part III of this Study.

B. TECHNICAL PROTECTION MEASURES (TPMS)

Occasionally, technical protection measures are attached to databases by copyright owners. Those TPMs prevent or limit the access to the contents and the use of the database (for example they will prevent reproduction of texts or massive downloads from databases and websites).

These TPMs are subject to a legal protection and could constitute a legal impediment to text and data mining. The InfoSoc Directive provides for a protection against the circumvention of any effective technological measures (art. 6 (1) of the InfoSoc Dir.).

However, if there is a conflict between one of the exceptions of the InfoSoc Directive and a TPM, article 6 (4) of the Directive provides for a possible intervention mechanism in case right holders do not balance the alleged conflict with voluntary measures. TPMs therefore cannot prevent the working of the exceptions to copyright[303]. Should an exception be created for data analysis, the same mechanism should thus logically apply.

C. DIGITAL RIGHTS-MANAGEMENT INFORMATION (DRMS)

The owner of data or texts can further add digital rights-management information ("DRMs") to its content. Following the InfoSoc Directive, DRMs are any information provided by rightholders which identifies the work or other subject-matter referred to in the Directive or covered by the *sui generis* right, the author or any other rightholder, or information about the terms and conditions of use of the work or other subject-matter, and any numbers or codes that represent such information (art. 7(2) of the InfoSoc Directive).

The InfoSoc Directive protects DRMs against their removal or alteration and the distribution, importation for distribution, broadcasting, communication or making available to the public of works or other subject-matter protected, from which DRM has been removed or altered without authority, if the person performing knows or has reasonable grounds to know, that by so doing he is inducing, enabling, facilitating or concealing an infringement of any copyright or any rights related to copyright as provided by law, or of the sui generis right (art. 7 (1) of the InfoSoc Directive).

It could be that, for some TDM projects, the works to be analyzed need to be separated from the DRMs identifying them, because, for example, all works need to be transformed in one and the same format or

[303] See on this issue the recent Nintendo decision of the ECJ: 23 January 2014, Case C-355/12.

in just a few compatible formats, while the DRMs are in other standards. Sometimes metadata will be useful, sometimes they may need to be removed.

Strictly speaking, it would seem that such removal (as the word is used in the InfoSoc Directive) of the DRM would be a violation of the protection of DRMs organized by the InfoSoc Directive. Yet, a teleological reading of this provision leads one to conclude that an infringement should only be considered to take place if the work is being reproduced and re-injected on the market and/or communicated without the DRM. An internal process (like a data analysis) which does not involve any distribution of copies of the work without the DRM or any communication or making available of the work does not harm the author. And in the output of the data analysis, as explained before, the works or data which have been used in the data analysis will normally not be copied: out of thousands of works, articles, databases, etc., the output will have found new trends or patterns and the output (the report) will describe this, without reproducing the works themselves.

We would see two solutions to avoid an absurd situation where the removal of the DRM, where necessary for a TDM project, would render this project illegal:

- Either an explicit legal provision is added to the new TDM exception to that effect;
- Or, at least, the explanatory memorandum to a possible new legislation makes this clear.

We have not seen this issue being mentioned in the literature. From discussions we have had with some interested parties, it seems logical to consider it, even if it has not lead in practice to discussions between stakeholders.

D. DATA PROTECTION AND IMAGE RIGHTS

One of the elements of the legal context which can create certain obstacles to text and data mining is the personal data protection legislation. Another (partly related) one comes from the rules on image rights.

The **Data Protection Directive**[1] applies to the processing of personal data. TDM unavoidably involves acts of "processing". [2] Reuse of personal data (e.g. "*a name, a photo, an email address, bank details, posts on social networking websites, medical information, or a computer IP address*."[3]) for TDM purposes would thus amount to a "processing of personal data". Personal data can only be gathered legally under strict conditions, for a legitimate purpose. Persons or organizations which collect and manage personal data must protect it from misuse and must respect certain rights of the data owners.

Again, by saying this, we are not suggesting that these rules should be adapted to facilitate TDM on personal data.

The reuse of a person's image during a mining process could constitute an infringement to the **image rights** existing in some of the selected Member States. The image rights are protected by article 8 of the European Convention on Human Rights and, to a certain extent, by the Data Protection Directive. In the selected Member States, the image rights are also sometimes further protected by criminal provisions (France), the Civil Code (specific provisions related to the right of personality: France and Hungary), press law (Poland), Copyright Law (Belgium) and principles of common law (in the UK).

[1] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, *O.J. L 281 , 23/11/199, p. 31 – 50.*

[2] Because "processing" means operations "*such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction*" (art. 2 Data Protection Directive).

[3] European Commission, "Data protection reform: Frequently asked questions", MEMO/12/41, 25 January 2012, http://europa.eu/rapid/press-release_MEMO-12-41_en.htm?locale=en

E. CONTRACT LAW AND LICENSE TERMS

The data processed in a TDM project may have been made available via a license agreement between the owner of the data and the "miner"; in other cases, the data will sometimes be accessible online on a website but the website will display terms and conditions which allow or, on the contrary, restrict (or even prohibit) the reuse of the data (and the mining of data).

Open access licenses, like Creative Commons, facilitate the use of data and thus also the mining of data (see elsewhere in this Study), so that these terms and conditions will not be an obstacle to TDM. However, in other cases, the owner of the data (the licensor) may only grant conditional access to the users, including a prohibition (or restrictions) to do any data mining on the data, for example through the terms of use of a website: the user will first have to accept these terms and conditions (by ticking an "I agree" box). Licensors do thereby control the access to their data and their databases via licenses.

The question as to whether these terms and conditions are enforceable upon users if the users were not obliged to explicitly agree (by ticking an "I agree" box) is uncertain: if the terms and conditions prohibiting or restricting uses of the data (e.g. prohibiting datamining on the data) were simply available on the website (be it in a very visible way or only after the user has clicked on an hyperlink to then access the terms and conditions), but the user did not have to manifest his/her consent to said terms and conditions, it is debatable whether these terms and conditions do apply (the answer to this question may vary from one Member State to the other and also very dependent upon the factual circumstances of the case (like: how visible were the terms and conditions?).

However, if (as we suggest it in our recommendations – see Part IX), a new specific datamining exception is introduced and considered unwaivable, then contract terms, license terms (included in DRMs information, in websites' terms and conditions or in negotiated license contracts) could no more impact the possibility to do TDM: the licensor would not have the possibility, via contractual terms, to prohibit datamining – at least as long as the other conditions of the exception would be respected[304]).

If this recommendation is not followed, and if the licensor remains free to prohibit datamining as he wishes, then obviously, contract law and license terms will remain a significant obstacle to datamining, since researchers will need to have the authorization of the licensor before they can start a TDM project.

F. LAWS ON UNFAIR/PARASITIC COMPETITION

EU Member States must regulate unfair business practices harming consumers' economic interests in accordance with the principles laid down in the Unfair Commercial Practices Directive [305] but unfair trade practices amongst competitors (B2B) is still largely not harmonized. In some Member States (notably, France and, to a lesser extent, Belgium), companies can complain in court that a competitor has committed **parasitic competition** by making profit of a competitor's efforts and investments without himself doing any effort and by putting a product on the market which he would never have been able to produce, should he not have acted in such "parasitic" manner. The theory is very much based on case-law and is not unanimously accepted, notably because it can be claimed that IP rights are in essence an exception to the freedom to copy, so that when an object is not protected by any existing intellectual property right, it can be freely copied (as long as no confusion is created and no illicit misappropriation took place, via theft or hacking of IT systems, for instance).

[304] i.e., in our recommendations, if it is (amongst other conditions) limited to scientific research and non-commercial purposes.

[305] Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) n°2006/2004 of the European Parliament and of the Council, OJEU, L 149/22, 11 June 2005.

Apart from so-called “parasitic behaviours”, unfair competition includes **misappropriation of confidential information**.

Under certain conditions, information contained in the data or text can be considered as confidential. Data mining could thus constitute an infringement to the protection of confidential information (called in legal terminology “trade secrets”, “undisclosed information”, “business confidential information”, “secret know-how”, etc.).

The definition of confidential information varies from Member State to Member State. Some general conditions of trade secrets are described in art. 39 of the Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS Agreement):

- The information is secret (in the sense that it is not generally known among or readily accessible to persons within the circles that normally deal with the kind of information in question);
- It has commercial value because it is secret; and
- It has been subject to reasonable steps under the circumstances, by the person lawfully in control of the information, to keep it secret.

G. SECURITY, SECRECY, UNAUTHORIZED ACCESS TO IT SYSTEMS

We can imagine that rules which protect the security of a State or of persons may sometimes also affect data mining. For example, anti-terrorism legislation, top-secret defense legislation, diplomatic secret, secret of the information services of the police, etc.

The access to governmental or military documents may be restricted to certain persons for reasons of national or international security.

Legislations on unauthorized access to computer systems (hacking) could also render some TDM projects illegal.

In the same sense, legislation on professional secrecy may restrict the access to certain data and constitute an impediment to TDM, for example banking or medical secret.

VII. POSSIBLE INITIATIVES WITHOUT LEGISLATIVE CHANGES

Apart from possible legislative changes (which will be examined under Parts VIII and IX), one may wonder if initiatives from the Commission which would not imply legislative changes may be desirable and possible regarding TDM. We suggest envisaging the two following options: (A) facilitate MoUs or other arrangements between stakeholders, and/or (B) adopting an interpretative document.

A. FACILITATING MOUS OR OTHER ARRANGEMENTS BETWEEN STAKEHOLDERS

This is an option which is probably always useful. This is already what the Commission has done by launching the Licenses for Europe exercise during the year 2013, with one workshop covering TDM. Participants did not all agree with the process and the TDM workshop witnessed some difficult debates, yet a "pledge" eventually came out of it and was presented, amongst 9 other "pledges", at the plenary session of Licenses for Europe. The pledge contains commitments by a series of scientific publishers to facilitate text and data mining of subscription-based material for non-commercial researchers. The pledge also complements this commitment by a standard license clause for subscription-based material, and announces future technological solutions to enable mining by researchers, at no additional cost, in journals subscribed by their university or research institutions (via a "mining portal", accessed after having accepted a click-through license).

However, solutions presented by publishers in Licences for Europe only cover part of the issues surrounding text and data mining and it remains to be seen whether they will be considered a viable solution for users (in particular the researchers community has often stated that they consider legislative changes as the only viable option way forward). Apart from contents available in subscription-based materials (periodicals, databases, etc.), there is other contents likely to present an interest for data miners: the Web in general, public domain sources, information held by the public sector (see our section on the PSI Directives), social networks, non-scientific materials, etc. They are not covered by the "pledge". Voluntary arrangements are to be welcomed but do not offer a global solution to the issues raised by data mining as a new phenomenon.

This option is therefore certainly welcome and should be encouraged but it does not appear to us as the only possible option nor as a sufficient solution for change.

B. ADOPTING AN INTERPRETATIVE DOCUMENT

Should legislative changes (via a new separate directive dealing specifically with TDM, or a directive amending the InfoSoc Directive and the Database Directive on various issues) appear not to be a feasible option, then at least an interpretative document from the Commission would be useful to clarify some issues on which there still is some debate:

- on the applicability of the "illustration" condition to the "scientific research" exception (to clarify that illustration only applies to teaching but not to scientific research;
- on some other issues where guidelines may be useful (as to what is "scientific research", what is a commercial or a non-commercial purpose, etc.).

Instead of waiting for the Court of Justice to further harmonize the exceptions to copyright set out by the Directives, piece by piece, somewhat haphazardly as the cases come up, the Commission could choose to draft and publish its own interpretative guidelines, not only on existing exceptions but possibly also on the freedom which users have, based on general copyright principles, to do TDM.

The Commission has already taken such initiatives in the past and in various fields, in order to clarify the scope and the meaning of the European legislation[306].

In the field of copyright, WIPO has also published its interpretative guide to the Berne Convention[307]. This guide was drafted by civil servants of WIPO, in order to clarify the provisions of the Berne Convention. The ECJ has already referred to this guide, even if it is not legally binding, while interpreting the provisions of the Berne Convention in copyright cases.

Guidelines aim at clarifying the commented instruments, in the light of the available case-law, and at providing the author's or institution's view about their interpretation. They should, ideally, clearly specify that they do not provide an authentic interpretation of the commented instrument. Indeed, when EU directives are concerned, the ECJ has the exclusive power to give such authentic interpretation.

At the same time, an interpretative "soft-law" instrument certainly has downsides. Its legitimacy may sometimes be questioned as it arises neither from the legislator nor from the Court of Justice which is entrusted by the EU Treaties to interpret Community legislation.

[306] See for instance the Guidelines on the interpretation of key provisions of Directive 2008/98/EC on waste published in June 2012 http://ec.europa.eu/environment/waste/framework/pdf/guidance_doc.pdf; the Commission Interpretative Communication on the Community law applicable to contract awards not or not fully subject to the provisions of the Public Procurement Directives of 2006 http://ec.europa.eu/internal_market/publicprocurement/docs/keydocs/communication_en.pdf and the Commission Interpretative Communication on certain aspects of the provisions on televised advertising in the "Television without frontiers" Directive, published in 2004 http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52004XC0428%2801%29:en:HTML

[307] Guide to the Berne Convention on the protection of literary and artistic works, published by WIPO, 1978.

VIII. WHAT LEGISLATIVE CHANGES COULD BE ENVISAGED?

A. INTRODUCTORY REMARKS

In this last part of the Study, we will, as per the Terms of Reference, "*assess the need, and if so the possible options, for legislative changes including whether there is a need to establish a specific exception to copyright for text and data mining activities (and, if relevant, to the* sui generis *right for databases*)".

More factual information/data needed. Based on the information available it is currently difficult to have a clear view as to what extent it is at the moment difficult for users to obtain the authorizations from the rightholders (i.e. mainly, the publishers) to engage in TDM. Statements made on this issue by the research sector and by publishers vary and partly contradict each other[308], and we have no factual information or market analysis which can invalidate either position. However, data analysis does not only take place on the basis on subscription-based contents obtained from publishers but may also take as research materials information which is available on the Web or in written form (sometimes being in the public domain), etc. The (real or alleged?) difficulties to negotiate licenses between research institutions/companies and licensors/publishers are not the only reason which would or should justify the adoption of an exception.

It is clear that text and data mining/data analysis is still a new phenomenon and that more factual and market information is needed. The public consultation launched by the European Commission in December 2013 might already provide some additional insight. Therefore, our suggestion would be that more investigation should be made (via a market and economic analysis).

What follows starts from the assumption that the present situation does not correspond to the best possible scenario for both researchers doing data analysis and for rightholders and that some improvements could be brought to the actual legislative framework.

We understand that researchers (from the academic sector or from commercial sectors) do not know exactly what they can or cannot do; as such, this insecurity may already be a good reason to clarify the situation.

B. SCOPE OF OUR SUGGESTIONS

Our recommendations are meant for TDM, not for scientific research in general. Please note that what will be suggested hereafter was only considered for the data analysis sector. Some of the suggestions we make for TDM may be relevant for the scientific research sector in general, and not just for the TDM sector, some other of our suggestions may on the contrary not be relevant or even desirable for scientific research in general. Our suggestions are however limited to the field of TDM.

A clearer legal framework is desirable. As we illustrated in the previous parts of this Study, a large part of data analysis is already covered by an existing exception, i.e. the scientific research exception existing both in the InfoSoc and in the Database Directive. However, we also indicated that the present situation creates some legal uncertainty, is not harmonized throughout the European Union and that some conditions of the research exceptions do not fit very well with data analysis. This is why we will, in the remaining part of this Study, suggest a specific TDM exception, which is inspired from the scientific research exceptions but contains a few variations compared to the scientific research exception in the InfoSoc Directive and in the Database Directive.

[308] See the discussions during the Licenses for Europe workshops related to TDM in 2013.

The need for a specific TDM exception follows, in our view, from the various issues that we identified in the previous parts of this Study and which either lead to legal uncertainty (which would thus gain to be clarified) or which raises problems or unjustified obstacles for TDM.

We will illustrate this in more detail hereafter and in Part IX.

C. A NEW EXCEPTION, SPECIFIC FOR DATA ANALYSIS

We will hereafter explain why we believe that an exception covering data analysis would be welcome.

Our suggestion is to have an exception which would be inspired from, and contain partly the same conditions than, the **scientific research exceptions**, but which would have its own characteristics. The suggestions we are making hereunder are made only for TDM, not for scientific research in general.

The main reasons which, in our view, justify that a specific TDM exception is added to the existing legal framework have been explained in the previous parts of this Study and, in summary, are the following:

- The present situation regarding the scientific research exceptions in the InfoSoc Directive and in Database Directive is not harmonized; for both Directives, Member States were free to implement the exception or not and, as we have seen, have not always done so, or not done so in a similar manner; this leads to a non-harmonized situation;
- The exceptions for scientific research can be waived by contract;
- There subsists a controversy as to whether the exceptions apply to "illustrate scientific research" or "for scientific research as such";
- The copyright exceptions are limited by the condition of "solely for scientific research", which excludes projects where, apart from a scientific research objective, there may be other ancillary objectives;
- The two Directives impose (with some variations) mentioning the authors' names and/or the sources, which does not really make sense for data analysis.

On the other hand, the fact that the exceptions are limited to non-commercial purposes should, in our view, not be changed.

What would be the relation between the general scientific research exception(s) in the InfoSoc Directive and in the Database Directive and the new specific TDM exception?

The specific exception we are suggesting would have its own conditions of application; the latter would be partly similar to the conditions contained in the other scientific research exceptions and partly different.

The specific TDM exception would thus not be subordinate or incorporated in the other scientific research exceptions: it would be a new and separate exception, to be added to the list of exceptions which already exist. The specific TDM exception would not be a *lex specialis* vis-à-vis the other scientific exceptions: in our proposal, the different exceptions are not in a hierarchical relation, they each have their own conditions of application and co-exist in parallel. There may be overlaps between the specific TDM exception and the existing scientific research exceptions (as there already are certain overlaps between some of the existing exceptions in the list of the InfoSoc Directive); a user may have the possibility to invoke one or more of the exceptions, if their respective conditions of application are present.

Whether the variations that we suggest for the specific TDM exception from the existing scientific research exceptions should also be implemented for the latter (and lead to a modification of the drafting of these existing scientific research exceptions) has not been considered here and was not the object of this Study.

IX. THE ELEMENTS OF A NEW EXCEPTION

We shall, in this Part IX, describe the elements of the new exception we suggest to adopt.

In some situations today, there is no need to invoke an exception (*reminder*). As indicated earlier, it should be reminded that, in the context of the present legal framework, some data analysis projects do not conflict with any copyright or any database rights, for instance, if the works used are in the public domain or if only insubstantial parts are used from databases.

In such situations, no exception needs to be invoked because no infringement happens in the first place.

As a consequence, the clause providing for a specific TDM exception should start by stating this, for instance in the following manner:[309]

> *"To the extent that data analysis involves works or data protected by the exclusive rights established in the [InfoSoc Directive] or the [Database Directive], the Member States shall introduce the following exception: (…)"*

Definition of TDM (*reminder*). We suggest that the new exception start by defining what "data analysis"[310] is, and we refer here to the definition we suggested at the beginning of our Study, i.e.:

> *"The automated processing of digital material, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights."*

The scope of the exception we suggest is limited to such data analysis.

a) A new exception inspired from the scientific research exception(s)

For the reasons explained above, we suggest that a specific TDM exception be adopted, along the lines of the scientific research exceptions in the InfoSoc Directive and in the Database Directive but with some variations. Our suggestion and the reasons for it will, where necessary, be described or reminded hereafter.

In certain – rather rare – cases, data analysis might be considered as falling within the **temporary copy exception** of article 5.1. of the InfoSoc Directive. As opposed to the scientific research exception in article 5.3. of the InfoSoc Directive, the temporary copy exception has the advantage, from a harmonization point of view, that it is a (the only) compulsory exception which Member States were obliged to implement in their national legislation.

However, as we have indicated above, the temporary copy exception is in our view not a solid basis to build on when trying to introduce a TDM exception:

- Its *rationale* is completely different and only in rare cases could a TDM project fall under this exception (mainly because it usually requires more than a purely transient copy): so, enlarging this exception to cover most other data analysis projects would either require a drastic change to this exception (i.e. abandoning the "transient" condition); this would completely change the philosophy of the exception and not correspond to its initial *rationale*;

[309] We were not asked to suggest a draft clause for the exception. So the wording we suggest sometimes here should not be considered, as far as its drafting is concerned, as a starting point for a future initiative. We only tried to give the idea which would need to then be translated into more precise and rigorous wording.

[310] As indicated previously, in a legal provision, we suggest referring to "data analysis" rather than to text and data mining. The other legislators who have adopted or intend to adopt new legislation in this section never used "text and data mining" as the key-word.

- Secondly, one knows that each and every word of the temporary copy exception has been the subject of heavy and difficult debates: changing it would re-open difficult debates and any change to the equilibrium which was eventually found in this article will unavoidably be jeopardized if a word of it is changed. Changes to its wording will most certainly open the way for applications of the exceptions to situations which were not intended to and not thought of by the lawmaker.

This is why we suggest to rather considering changes inspired from the **scientific research exceptions**, along the lines described hereafter.

The purpose (and at the same time, the scope) of the new exception would be **scientific research**, as such terms are already understood in the InfoSoc Directive and in the Database Directive. We do not suggest deviating from this concept for TDM.

The word "scientific" must continue to qualify research and should not be abandoned. Since we suggest using the scientific research exception(s) as a source of inspiration for a specific TDM exception, we will hereafter describe which other elements from these exceptions could be incorporated (with or without changes) into the specific TDM exception.

The scientific research exceptions are provided for in article 5.3.a of the InfoSoc Directive, article 6.2.b of the Database Directive (copyright) and article 9.b of the Database Directive (sui generis right).

These articles state resp. as follows (the parts which are relevant for our Study are here emphasized):

Article 5.3.a., InfoSoc Directive:

"Member States may provide for exceptions or limitations to the rights provided for in Articles 2 and 3 in the following cases:

(a) 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****";*

Article 6(2)(b) of the Database Directive (copyright):

"Member States shall have the option of providing for limitations on the rights set out in Article 5 in the following cases: […]

(b) 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".***

Article 9(b) of the Database Directive (sui generis right):

"Member States may stipulate that ***lawful users of a database which is made available to the public in whatever manner*** *may, without the authorization of its maker****, extract or re-utilize a substantial part of its contents****:*

[…] (b) in the case of extraction ***for the purposes*** *of illustration for teaching* ***or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved****".*

As one can see, there are some differences in the wording:

- mentioning **not only the source but also the author's name** in the InfoSoc Directive;

- **unless, however, this turns out to be impossible** (such safeguard is not provided in the Database Directive, neither for copyright nor for the sui generis right);
- **sole purpose** of scientific research regarding copyright in both the InfoSoc Directive and the Database Directive, but "**purpose of** scientific research" for the sui generis right.

One condition is consistently included: "**to the extent justified by the non-commercial purpose to be achieved".**

Our proposal would be the following (and we will further give the reasons for this hereunder):

- the condition which is consistently used in all three provisions should be kept : "**to the extent justified by the non-commercial purpose to be achieved";**
- the "sole purpose of scientific research" should be replaced by "**having as its main purpose scientific research**" (this has already been explained above);
- the **mentioning of the author's name and of the source**, regardless of whether there is an escape clause (as in the InfoSoc Directive) or not, should be abandoned and left at the discretion of the data miners.

Why do we or, for one condition, do we not, suggest these changes?

(i) Non-commercial purpose to be achieved (no change)

The criterion **"justified by the non-commercial purpose"** should remain, for reasons we have partly already mentioned when considering the different exceptions on scientific research and which we will hereunder further elaborate: we see no compelling reasons why information providers or publishers, who are making content available to researchers/companies in order for them to do TDM, if this takes place for commercial purposes on the side of the users, would not have the right to benefit from exclusive rights and from the right to claim royalties for such commercial use.

It is commonly admitted that "commercial" should be read as including direct and indirect economic and commercial advantages. It has been stated on various occasions that the commercial v. non-commercial criterion is difficult to apply. This may be true and obviously there are borderline cases, but this is in itself not an argument for abandoning the criterion. Even if there exist some borderline cases, in the majority of cases, the answer as to whether a purpose is commercial or not will be quite clear or totally clear. One does not abandon a rule because it is difficult to apply it in a minority of cases while it remains an efficient tool in the majority of situations.

The exclusion of commercial purposes from the benefit of the exception is however to some (growing) extent mitigated by the development of Open Access licenses and Creative Commons licenses: not all of them but some of them do allow reuse of the works (scientific works for Open Access) *also for* commercial purposes. This, together with the impact of legislation on the re-use of public sector information (PSI) is a growing trend and it does open new possibilities for data analysis for commercial purposes (see other part of this Study on this).

One may notice that the UK proposal for new legislation mentioned at the beginning of this Study also only aims at including under the new suggested data analysis exception *"the sole purpose of non-commercial research".*

As long as the exception remains limited to non-commercial purposes, it means that all commercial companies need to acquire licenses from the publishers. The argument has sometimes been put forward by some that it would mean that only very large commercial companies could afford paying the price for such licenses, which would thus be limited to Internet giants or to multinational pharmaceutical companies while SMEs could not afford such licenses. This does not seem realistic to us at all: publishers will much like software publishers, obviously apply different tariffs depending upon who the user is and will normally not charge the same fees to a large multinational search-engine than to a local startup company embarking in a new risky endeavour. The argument therefore seems ill-founded;

If we were to expand a TDM exception to scientific research even for commercial purposes, this would mean that companies would not have to pay anything to rightholders to do data analysis on the contents they acquired, while at the same time, such analysis does require to perform acts falling under copyright exclusive rights (as explained before in our Study). This does not seem like a logical solution to us: these companies, much like small companies, have to pay all their software, hardware, research equipment, office appliances, etc. at normal prices and cannot claim that, just because they do scientific research, they are entitled to obtain all these materials and software for free or at preferential conditions. This should be all the more true when it concerns the acquisition of scientific information or data, or published articles, which will serve as indispensable raw material to their core business of research. It may be that the companies doing scientific research already obtain tax deduction advantages (which may be justified to promote research) and maybe other advantages like lower social charges when hiring researchers (which may also be justified for the same reasons) but the costs of these advantages are borne by the State; we fail to see good reasons why they would be entitled to obtain preferential treatments from private publishers and why publishers (particularly STM publishers) would have to grant them preferential treatments or tolerate exceptions; this would deprive them of the possibility to recuperate the investments they made to facilitate access to their clients to subscriptions in digital formats that are more easily usable for TDM.

The question may further be raised whether an exception granted also for commercial purposes would pass the three-step test.

A possible exception for text and data mining should comply with the three-step test, described in article 5(5) of the Directive and in various international instruments[311].

It is hard to respond to this without more factual information, and such assessment always relies on a prior large market investigation. The analysis would need to be substantiated with objective economic and market figures, which we do not have. Absent such information, we could only make a number of observations:

1. The limitation should be restricted to "certain special cases" (step 1)

The exception or limitation should be clearly defined and it should have a narrow scope (in a quantitative sense)[312].

The legislature can provide an exception or a limitation to the exclusive rights that pursue particular objectives (favouring public interest objectives and in order to reconcile the protection under copyright with other causes, such as freedom of expression and information, education and research). According to the WTO panel, the justification of the exception in terms of a legitimate public policy purpose should not be verified under the first step[313] (this may be verified under the second and third steps[314]).

[311] Art. 9(2) of the Berne Convention, art.10 of the WIPO Copyright Treaty and art.13 TRIPS Agreement. The ECJ declared that the provisions of the TRIPS Agreement are applicable in the legal order of the European Union, hence also the TRIPS three-step test ECJ, 15 March 2012, C-135/10, *SCF v. Del Corso*, para. 56.

[312] WTO, Report of the panel, 15 June 2000, WT/DS160/R, United States, Section 110(5) of the US Copyright Act, available at http://www.wto.org/english/tratop_e/dispu_e/1234da.pdf. See S. RICKETSON & J.C. GINSBURG, *International Copyright and Neighboring Rights*, Oxford University Press, 2006, 764; M. SENFTLEBEN, *Copyright, limitations and the three-step test. An analysis of the three-step test in international and EC Copyright law*, Kluwer International, The Hague, London, New York, 2004; S. DUSOLLIER, « L'encadrement des exceptions au droit d'auteur par le test des trois étapes », *I.R.D.I.*, 2005, p. 213-223.

[313] WTO Panel Report, 33, par. 6.111.

[314] J.C. GINSBURG, "Towards Supranational Copyright Law? The WTO Panel: Decision and the "Three-Step Test" for Copyright Exceptions", January 2001, Revue Internationale du Droit d'Auteur, January 2001. Available at SSRN: http://ssrn.com/abstract=253867. See also RICKETSON & GINSBURG, *International Copyright and Neighboring Rights*, 767.

Enlarging the benefit of the exception for commercial purposes could only raise doubts regarding the "special cases" condition: if the exception covers both non-commercial and commercial purposes, it could be that, already from a quantitative point of view, the exception is too broadly conceived, as it would cover all purposes for which TDM can be performed and potentially benefit all potential users of TDM. It would then not be limited to "special cases" and would not pass the first step.

2. The exempted use does not conflict with the normal exploitation of the work (step 2)

The exclusive rights under copyright can be understood as protecting the exploitation of a work, the faculty of an author to realise economic value from the use of the work. Any exception may have an incidence on the exploitation of the work but only if it conflicts with the *normal* exploitation should it be deemed unacceptable. The "normal exploitation" refers to the markets that the author is already addressing, those that he is likely to develop in the future and the potential forms of exploitation that "with a certain degree of likelihood and plausibility, could acquire considerable economic or practical importance"[315]. Many commentators present data analysis as an important future economic sector which should be nurtured and promoted and of which the economic importance will grow. This means that this mode of exploitation of works and data is likely to grow and become of great importance for publishers and rightholders. For the STM publishers, an important distribution channel of their scientific journals is obviously the companies doing research. Data mining is probably due to become a primary mode of exploitation of scientific publications.

The second step also requires an economic analysis of the commercial use of the exclusive copyright rights, and "not every use of a work, which in principle is covered by the scope of exclusive rights and involves commercial gain, necessarily conflicts with a normal exploitation of that work" [316]. An exempted use may conflict with the normal exploitation of the work if it "[enters] into economic competition with the ways that right holders normally extract economic value from that right to the work (i.e., the copyright) and thereby [deprives] them of significant or tangible commercial gains" [317].

It has however been argued that this second step also calls for a non-economic normative consideration, in particular the question whether the author should be entitled to reserve a market that may be affected by an exception favouring public policy objectives, such as education, research or library uses[318]. RICKETSON and GINSBURG thus invite to consider both economic and non-economic arguments and to proceed to a balancing of such considerations. A value judgment is required and the justification of the exception will need a "clear public-interest character that goes beyond the purely individual interests of copyright users" [319].

This second step demands an economic analysis of the "markets" for authors, which may be affected by the exception for text and data mining. It also invites to an examination of prospective markets and forms of use, protected under copyright's exclusive rights. This may be more challenging. Insofar as this is not yet the case, data mining may develop into an autonomous form of exploitation for (large) publishers of scientific journals or databases.

An exception that allows anyone to reproduce protected elements in view of performing data analysis for scientific research but with commercial purposes may undermine such exploitation of the work based on these (emerging) technologies. This is e.g. the case for TDM done by commercial intermediaries that offer TDM services with a commercial intent and in exchange for fees. Such services may prove the existence of a market for TDM services based on acts protected under copyright; an exception that

[315] WTO Panel Report, 48, par. 6.180.

[316] WTO Panel Report, 48, par. 6.182.

[317] WTO Panel Report, 48, par. 6.183.

[318] See on this point RICKETSON & GINSBURG, *International Copyright and Neighboring Rights*, 771.

[319] RICKETSON & GINSBURG, *International Copyright and Neighboring Rights*, 773.

exempts all protected acts on condition that the data analysis is done for the purpose of scientific research would give a free pass to such commercial entities, as long as their customers use the resulting data in the framework of their research (even if their customers may be scientific institutions or academic research centres without commercial intent). Similarly, companies that engage in research in order to develop products (such as medicines) for commercial exploitation would thus be exempted from obtaining the right holders' authorisation and thus deprive them from possibly important fees: if such commercial research entities are covered under an exception, the right holder has no ways left to benefit from this new form of exploitation of her work. Data analysis is in practice in many cases done for scientific research purposes.

Depending on the public policy objective that the legislature pursues[320], this condition could be understood to cover only academic research (at universities or publicly funded research institutions) or also collaborations between such "public" institutions and private companies. Only an economic analysis of the impact of this new form of exploitation would allow the legislature to better assess whether the second step is met.

3. The use does not unreasonably prejudice the legitimate interests of the rightholder (step 3)

An exception that meets the second step should not inflict unreasonable prejudice upon the legitimate interests of the rightholder. The *legitimate interests* of the author/rightholder are thus further protected, pecuniary and non-pecuniary interests that are pursued for considerations within the "proper" sphere of copyright[321]. These legitimate interests should not suffer unreasonable prejudice due to the application of the exception. This may sometimes be avoided by the payment of remuneration, as a compensation for the exempted use and the income that the author (or holder of other rights) may be missing. In this regard, M. Walters and S. Von Lewinski explain (in relation to another exception but the reasoning could be enlarged to commercial uses) that:

> *"[…] The Directive does not require that the right holders receive fair compensation. However, the application of the three-step test under Article 5(5) of the Infosoc Directive may result in an obligation of the Member States to provide for some form of fair compensation or remuneration, depending on the individual exception or limitation under national law.*[322]

Some have thus expressed the view that TDM should be authorised also for commercial purposes, but accompanied with a system of an equitable remuneration. This would in practice mean that authorisations would be granted not by the publishers or rightholders but by collective rights management organisations. We are not convinced that a mandatory collective solution is needed, because it has not been proven that it is impossible to negotiate licenses with STM publishers.

Obviously, where a TDM project does not aim at covering scientific periodicals but e.g. content available on social networks or on hundreds of websites, it does indeed become very difficult to ask for the authorisation (provided that such authorisation is needed because reproduction of copyright protected material is involved). But if a collective organisation is set up to grant the authorisations, how would it redistribute the fees to the myriads of rightholders who, individually, would never have complained about the fact that one of their Tweets, one sentence of their blog or of their website has been copied ? The solution should probably not be as easily discarded as this Study seems to do it, but we tend to think that

---

[320] The second step should indeed not prevent the exception from attaining the public interest underlying it. See on this point J. GRIFFITHS, "The 'Three-Step Test' in European Copyright Law - Problems and Solutions" (September 22, 2009). Queen Mary School of Law Legal Studies Research Paper No. 31/2009. Available at SSRN: http://ssrn.com/abstract=1476968, 11.

[321] RICKETSON & GINSBURG, *International Copyright and Neighboring Rights*, 774.

[322] "As an example of such remuneration in the case of reproduction of works for textbooks, see § 46(1) and (4) of the German Copyright Act, which already existed before the adoption of the Infosoc Directive". WALTER, M. W., and VON LEWINSKI, S. V., European Copyright Law op. cit., note 1, pp. 1045.

a mandatory collective solution with an equitable remuneration system would be difficult to implement and not likely to benefit rightholders from whom just very little pieces of work have been copied.

One could also sustain that collective management options are an acceptable solution only if the market cannot itself offer solutions based on individual license schemes. We do not have sufficient evidence to say that this is not the case. Our ***a priori* conclusion/impression** would be that an exception encompassing commercial purposes would run contrary to the three steps test but we should admit that we do not have the factual evidence and the economic data to make a substantiated statement on this difficult issue.

(ii) Not "solely" for scientific research

In the new TDM specific exception, we would suggest not to use the words "solely for scientific research" (from the scientific exception in the InfoSoc Directive) but rather to have as a condition that the data are used "**mainly for scientific research**".

"Mainly for scientific research" means:

i. that the purpose of scientific research must be **present**;
ii. that this purpose of scientific research must be the **principal** objective;
iii. but that there may be other purposes pursued **in addition**: the existence of other purposes does not exclude the application of the exception.

"Mainly for scientific research" does not mean any dilution of the scientific research exception; in other words, it does not mean that there should just be "some kind of scientific research" or even just "some kind of research". Also, "mainly for scientific research" does not mean that commercial purposes could suddenly be pursued (we suggest keeping, as for the other scientific research exceptions, the non-commercial purpose condition – *cf. supra*).

But this wording leaves some room for projects that may, besides scientific research, accidentally or incidentally pursue other purposes, without those purposes being the main objective. The expression "solely for scientific research" does, at least in principle, exclude such projects from the benefit of the exception, while in our view the mere fact that, besides a scientific research purpose, a project may incidentally have other purposes, should not deprive such project from the benefit of the exception. If the *rationale* underlying the exception is that scientific research benefits society, the exception should logically, in our view, not only cover projects having exclusively a scientific research purpose but also projects which aim "mainly" at scientific research.

The purpose of data analysis is often described as being to uncover new knowledge or insights from previously known data. It must be verified, on a case-by-case basis, whether a given data analysis project adds something to the state of science in order to be qualified as scientific research. In most cases, we would think that it will; in those cases, the exception will benefit them.

The word "mainly" might lead to somewhat more uncertainty and more borderline cases than the word "solely". In itself, this is not a sufficient reason not to adopt it; many of the existing conditions in the InfoSoc Directive or in some other directives also contain elements of legal uncertainty and this is to a large extent unavoidable.

Alternatively, if it is considered that "mainly" will lead to many uncertainties, one may decide that the same objectives could also be pursued – as an alternative - by simply dropping the word "solely" and keeping the words "for the purposes of scientific research", like in the Database Directive.

b) An exception not just to illustrate scientific research

In order to avoid the controversy which exists regarding the scientific research exception, it should be clarified that the purpose here is not to facilitate the "illustration of scientific research" but scientific research as such.

We do not think that the purpose of "illustration for teaching" in regard of TDM does correspond to a need. This may have to be further examined. Should it not be so, we would then suggest dropping the "illustration" condition in a possible new specific TDM exception.

c) No obligation to mention the authors' names and/or the sources

Mentioning of the authors' names and/or of the sources, even except if this proves impossible, does not offer any advantage to the authors/sources, since the output of the TDM project is a report where normally no excerpt from the pre-existing sources is being reproduced. Should it however be the case, then the quotation exception would provide for the conditions to mention the source or the author's name.

We already indicated that one may question the relevance of this obligation in the context of data analysis for the following reasons:

- this requirement seems logical for the exception regarding illustration for teaching because in that case, a work is being reproduced (or communicated to the public) and is thus "visible" in the teaching materials (so that every copy of such materials involve a copy of the pre-existing works): it is thus understandable that, in such cases (like for the quotation exception) the source, including the author's name, must be indicated;
- in many scientific research projects however (and particularly in TDM projects), the research project or process may necessitate to copy pre-existing works but it may well be (and, in TDM projects, it will almost always be the case) that the research output does not include any part of the pre-existing works but only draws conclusions (new insights, new patterns) thereon[323]. In such situations, a copy of the output (as opposed to a copy of the teaching materials) will not involve any copy of the preexisting works.

For the rest, the researchers will normally have kept the list of all the sources and authors but they will in principle not disclose them together with the TDM report/output: the report could be short, while the listing of thousands of sources and authors might be ridiculously much longer (with no one reading it). The choice of listing all their sources should thus be, like in the bibliography of a scientific publication, left at the discretion of the authors of the report. Peer pressure may push them into publishing the sources, and giving the necessary authoritative character to their report may push in the same direction.

But imposing it by law has several disadvantages: the burden may be high, while the advantage for the source or author to be mentioned in a list of thousands of other authors may be close to inexistent. And another concern, we think, should be taken into account: to a certain extent, disclosing all his sources may, for the researcher, disclose part of his know-how; further, it may be also that in some sensitive projects, having to disclose all his sources may expose him to aggressive reactions if the conclusions of the report are very unfavourable to some interested parties.

All in all, the burden is heavy, the advantages to the authors are close to zero. We suggest that this decision be left to the researchers; peer pressure or the need to convince their peers about the seriousness of their results may lead them to decide to list their sources.

We do not think that this proposal would be in violation of moral rights, in all cases where the output of the TDM (the report) does not even quote any of the pre-existing works. From a copyright standpoint, where

[323] See indeed our conclusion above in this Study and the difference we make between the data analysis process and the data analysis output.

no extract of the source materials is copied in the TDM output (the resulting report), one may argue that such mentioning of the source should not be made compulsory but left to the discretion of the researchers.

d) An exception to which exclusive rights?

We will here examine to which exclusive rights the exception should apply.

We suggest that the wording refers to the different exclusive rights under copyright and under the *sui generis* (database) right which we considered as being relevant. The principle would be that data analysis corresponding to the definition would not require the authorization of the rightholders of the works or data which would be analyzed.

Practically, the exception would be an exception to:

- the **reproduction right** in the **InfoSoc Directive**;
- the **reproduction right** in the copyright part of the **Database Directive**;
- the **adaptation right** in the copyright part of the **Database Directive**; more precisely, this should cover " *translation, adaptation, arrangement and any other alteration"* of the database[324]*;* and in order to be exhaustive, this should also include the reproduction of the results of the acts of "translation, adaptation, arrangement and any other alteration of the database"[325];
- and the **extraction right** in the *sui generis* part of the **Database Directive**.

As indicated previously, a TDM exception would arguably also be an exception to the adaptation right[326] and the translation right which are not harmonized in the InfoSoc Directive (the adaptation right is only harmonized in the Database Directive). For reasons we explained, this is in our view not an obstacle (in the same way as it has not been an obstacle to introduce e.g. the parody exception in the InfoSoc Directive and because these adaptations and translations do, for TDM, also qualify as reproductions under the InfoSoc Directive).

It is not our understanding that moral rights need to be mentioned and that exceptions to moral rights would here be necessary – see what we suggested under c) regarding the mentioning of the source.

e) A compulsory exception for Member States

As described in our general Study[327], the manner in which the exception for scientific research in Article 5.3 a) of the InfoSoc Directive was drafted leaves various questions unanswered. Furthermore, the implementation by the Member States has led to diverging legislations. What we said in our general Study about the utility of having more harmonisation for the exception for scientific research in general holds true, *mutatis mutandis*, for data analysis. The differences arising from the lack of harmonisation can only lead to complicated situations for cross-border projects, for scientific research in general, and for data analysis as well

The fact that most of the exceptions in the InfoSoc Directive were optional and that their adoption was left to the discretion of the Member States' legislators[328] has been criticized by many. The result is what it is,

[324] According to Article 5.b of the Database Directive.

[325] According to Article 5.e of the Database Directive

[326] The adaptation right is explicitly mentioned in the Japanese exception on data analysis – see our presentation of the Japanese legislation *supra*.

[327] TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", De Wolf & Partners, October 2013, p. 109 et s., available on http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf.

[328] See our reasoning on this same issue in the chapter on User Generated Content in our Study, TRIAILLE, J.-P. (ed.), "*Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive")*", De

i.e. a rather disharmonized picture, with every country having made its shopping-list in the catalogue. We do not think that the proportionality principle or the subsidiarity principle *per se* justify that each exception in the catalogue list remains optional for the Member States. Directives are *per se* instruments which leave some discretion to the Member States as to the means to be put in place to reach the objectives of the directive; but that does not mean that a directive should always only include "optional" lists. Quite on the contrary, if the Member States are left free to decide on the means, they are nevertheless under an obligation to meet the objectives of the directive. Making the exceptions (at least some of them) mandatory would only align the regime of the Directive with the regime of the greatest majority of harmonizing directives.

As we have illustrated in our main Study, the exception of scientific research has not been implemented in the same manner in all Member States. We only covered, as was asked by the Commission, 11 Member States, but other Member States, not within this list of Member States that we examined, also implemented the exception in yet different ways. Sometimes the beneficiaries differ from Member State to Member State, sometimes "research" is understood differently, sometimes only parts of works can be used and sometimes works in their entirety may, etc.

This leads to a situation where not all researchers and research institutions are on the same level playing field and to a situation in which, depending from where the works originate from, they may or may not be used in a TDM project. For cross-border projects (either involving researchers from several Member States or works whose rightholders are from different Member States), this makes the situation difficult to assess from a legal point of view. For the same reasons, it does not facilitate pan-European research projects, which runs contrary to some of the Commission's objectives and policies.

When translated into TDM, the logical remedy would be to make the implementation of the specific TDM exception **compulsory** for Member States.

f) An unwaivable exception

Another question could be whether parties should be free to discard the scientific research exception and more particularly, for the sake of our analysis, to prohibit data mining for scientific purposes.

In the first part of this Study, we differentiated between four types of access to data which we qualified as (i) "all to all" / (ii) "many to many" / (iii) "one to many" / (iv) "one to one".

In case (i), no terms and conditions apply; at least, no terms and conditions are accepted by users, even if they are accessible somewhere on the front page of the website. In case (ii), the conditions are accepted by the user when opening his account on the social network. In case (iii), contractual conditions are agreed to by the client. In case (iv), contractual conditions are often negotiated on a one-to-one basis before being agreed.

The importance of contract law, as a sort of "second layer of protection" in addition to copyright, is growing when going from (i) through to (iv).

We suggest making the exception unwaivable (in the situations of course where all the other conditions for the application of the exception are met, i.e. scientific research for non-commercial purposes, lawful access). This also would help promoting scientific research in Europe and provide an incentive to the data mining sector by ensuring a level-playing field throughout the European Union, thereby facilitating cross-border projects. We do recognize however that the economic consequences of making this exception unwaivable have not been examined in the scope of this Study (this was not within its scope). Such economic analysis should be done to assess its impact; the obvious advantage thereby granted to the research sector should not be disproportionate to the disadvantages it would probably bring about for scientific publishers: researchers need publishers and publishers need researchers, so the *quid pro quo*

Wolf & Partners, October 2013, p. 109 et s., available on http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf.

needs to be established with due care and prudence. We therefore recommend that the economic consequences of our recommendation be analysed.

The impact of this should be assessed differently by distinguishing between scientific publishers and other rightholders:

- For scientific publishers, licenses to do data analysis on their assets (scientific journals, scientific databases, etc.) are a potentially important and growing source of income; the impact of making the exception unwaivable risks to have a negative impact on this. For some other rightholders, data analysis is only an incidental use of their works/content; this would be the case for most non-scientific information, like novels, artistic works, information generally available on websites or social networks, etc.: all such elements may be useful in the context of certain TDM projects (for instance, linguistic research), but a writer of a novel or a blogger do not expect that their works will be mined and that they will be able to derive a benefit from this (unexpected) use.

However, we only suggest this for categories (ii) and (iii) above:

- For (i) (web data): there are no contractual terms imposed upon users, so there is no need to consider the issue of the unwaivability of the exception by contract[329];

- For (iv) (confidential data): the agreement is concluded on a one-to-one basis and the terms are often the subject of negotiations. Our suggestion cannot lead to the consequence that non-disclosure agreements (which are essential agreements in many situations, particularly in the research and innovative sectors of industry) would become unenforceable. Confidential data should continue to be governed essentially by the confidentiality agreements and the disclosing party should remain free to impose conditions including the prohibition of doing data analysis on such data. Most likely, as we deal here with unpublished data, the exceptions from the Infosoc Directive or of the Database Protection Directive do not apply anyway. It may be worth verifying whether the draft directive on the protection of trade secrets would hereby be impacted[330].

If the purpose is that the exception can always be relied upon by users, even when they are first invited to accept standard clickwrap terms and conditions, *then* a further logical step would be to decide that such exceptions be made "imperative" (as so called in civil law countries), or in other words unwaivable by contract[331]. Making some of the exceptions unwaivable by contract, particularly those which exist not due to some market failure but result from freedom of expression considerations or from public policy objectives (like the promotion of R&D activities in the European Union), may make sense.

Some countries like Belgium have already made copyright exceptions unwaivable by contract; it does not seem that this has led to dramatic consequences for rightholders. In consumer law legislation, many provisions protecting the weaker party have been made unwaivable[332]; it would seem that there are comparable reasons to adopt similar mechanisms in copyright law, for certain exceptions at least.

If one is of the opinion that making an exception unwaivable is too far-reaching or too simplistic, an alternative could be to distinguish between adhesion contracts and contracts that are negotiated (one could maybe add: "at arm's length"): exceptions would only be unwaivable in adhesion contracts and not in negotiated contracts.

[329] This assertion may need to be somewhat nuanced for there are cases where our conclusion is too simplistic. But studying this issue in detail should be done in the context of the assessment of the e-commerce directive and it here out of the scope of our study.

[330] Proposal of 28 November 2013 for a directive of the European Parliament and of the Council on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, COM/2013/0813 final, 2013/0402 (COD).

[331] This is the reasoning behind the legislation which, in a few European countries, have provided for such unwaibility of certain (or all) exceptions.

[332] In the field of copyright, the EU directive on the legal protection of computer programs, the back-up copy and the decompilation exception have been made unwaivable.

By suggesting this however, we are not suggesting that all exceptions to copyright should be made mandatory[333]. Some have clearly stated that a case-by-case (exception-by-exception) analysis should be made, before deciding, for each exception, whether it should be made unwaivable or not[334].

The main disadvantage is that parties lose all margin of negotiation, which may make certain transactions more difficult. It could lead to a situation where a rightholder, because he cannot reserve some uses, decides not to grant any license to the candidate user and that, as a consequence, some transactions which would otherwise have led to some exploitation of a work, are not concluded.

Another argument pleading for the unwaivability of the data analysis exception is the following:

It is a universally accepted principle that copyright protects the expression, not the ideas nor the information contained in the expression given to a particular work. There are some precedents where access to the ideas underlying a copyright protected work was barred due to a (too) strict application of the prohibition deriving from the reproduction right. Indeed, where copyright was first applied to computer programs, it has been accepted that the notion of reproduction in the Computer Program Directive was also very broad. Consequently:

(1) decompilation was becoming an infringement to this reproduction right and access to underlying information on interoperability of software with other software elements was becoming *per se* infringing;

(2) to observe, study or test the functioning of the program in order to determine the ideas and principles which underlie any element of the program would also have become infringing.

To avoid that protection of the form would bar access to underlying information, exceptions were introduced both for decompilation and for observation and study.

One should add that these exceptions were accompanied by a series of additional conditions before they could be relied upon, but the logic is the following: where the expression (protected by copyright) is an obstacle to accessing underlying information or ideas (unprotected by copyright), then a compromise needs to be found in the form of a conditional exception.

The analogy with text and data mining seems at least partly relevant: TDM intends to extract new ideas from protected works and, for technical reasons, in order to do so, reproductions need to take place before the analysis can be launched. But in all these cases, the purpose is not to copy the expression (i.e. the computer program or the works being analysed), it is to extract information from the data/works.

It is interesting to note that, in the Computer Program Directive, both the decompilation exception and the observation and testing exception were made unwaivable also (be it under certain conditions).

One may note that in the UK draft legislation on data analysis, it is also suggested that the exception be unwaivable.

It must however be made clear that the exception will only benefit users which had a lawful access to the data or works. In other words, making the exception unwaivable does not entail that any user would suddenly have the right to access any database and/or circumvent technical systems or "hack" the existing access protection for the mere reason that such user wishes to do a TDM analysis (see hereafter on "lawful access as a condition of the exception").

g) Lawful access as a condition of the exception

The question should be raised whether the exception should only benefit lawful users or users having a lawful access to the database.

[333] On this subject, see the PhD thesis by Lucie Guibault (IVIR, 2002), Copyright Limitations and Contracts: An Analysis of the Contractual Overridability of Limitations on Copyright.

[334] See e.g. Dr de WERRA, at www.ip.mpg.de/files/pdf1/TagungInteressenausgleich-2-Diskutant1.pdf

We tried earlier in this Study to illustrate that the "lawful use" terminology has lead to much confusion and that its interpretation leads to controversies and often ends up in "circular" argumentations:

- If it is only to be interpreted as meaning "users who can invoke the benefit of an exception", then we do not see what the added-value is: either he can benefit from an exception, and the user is a lawful user, or he cannot benefit from an exception and he is not a lawful user. We see no in-between situations;
- If it is to be interpreted as meaning users who respect the terms of use and the conditions of access to a database provided by the publisher (or the library), it means that said conditions of access can easily prohibit data analysis (and one knows that this happens): in that case, the exception is completely circumvented and becomes useless in practice;
- If, finally, one decides that the exception for datamining is unwaivable, then also we do not see what the added-value of the term "lawful use".

We do however not suggest that the exception should benefit unauthorized users who, for instance, have hacked security mechanisms put in place by a publisher (password, etc.).

Our suggestion would be the following:

- To avoid the confusion associated with the terms "lawful use" and "lawful user", one should rather refer to the expression "lawful access";
- Access should be considered lawful if it is either authorized by the rightholder *or* by an existing exception (e.g. the scientific research exception);
- It should be made clear that a researcher could not invoke the specific TDM exception to bypass technical protection measures put in place by the publisher or to require access;
- For the avoidance of doubt, it should be made clear that where data are available on a website without any restriction, data analysis made on such data is presumed to be made with lawful access;
- The provision would be without prejudice to the application of the provisions on TPMs from the InfoSoc Directive, including article 6.4.

For example, if archives of a newspaper are put on a website but are only accessible via a subscription (and the use of a password), the TDM exception would not benefit a user who accessed the system without the authorization of the publisher (e.g. by hacking the access code); such a user would be considered as not having a "lawful access". However, once a user would have obtained access to these archives (i.e. with the authorization of the publisher), the terms of the subscription agreement should not be allowed to prohibit data analysis on such archives (since we suggest that the exception could not be overridden by contract).

The UK draft proposal on data analysis does mention lawful access as a pre-condition to the suggested exception.

h) Non-substitutability as a condition of the exception

The question may be raised as to whether data analysis should be authorized even if the result of the analysis substitutes the original source, i.e. if makes it unnecessary to still access and read the original documents.

In the *quid pro quo* which would lie at the basis of this new exception, it may seem reasonable to impose that the output does not substitute the original sources.

In practice however, this will only rarely be the case. The output of a data analysis relies on hundreds or thousands of documents/data/pictures. It is unlikely that the output would substitute any of the original documents on which the analysis was based. The output will only extract very little information from each of these sources to compile them and try to draw new patterns from them. Each such source will, in addition to this little information which is being extracted for the sake of the data analysis, contain a lot of

other information and data which would not be relevant for the data analysis but may still be of interest for other purposes. Also, it may be that researchers interested in the output of the data analysis will, on the contrary, wish to consult the source documents (rather than abstain from doing so), to double check some information, links, etc. One should not start from the wrong assumption that the output of a data analysis would be a sort of summary of all the pre-existing materials.

As a consequence, this condition for the application of the exception would not often constitute an obstacle to said application. Adding it to the conditions of the exception will not offer much protection for the rightholders or publishers. In our opinion, the commercial v. non-commercial purpose is a much more important distinction that the (non-)substitutability.

The condition may be hard to apply and, as for other criteria in copyright law, there may be borderline cases. But this, as such, is not a sufficient reason not to introduce the condition. On a case-by-case basis, courts could be called upon to apply it and progressively clarify what the condition means in practice.

To be exhaustive, one should remind the Microfor/Le Monde case in France, from the French Supreme Court, where a summary of articles by Microfor was held to be infringing on the copyright of the newspaper Le Monde, for the reason (to summarize the reasoning here) that the summary proposed by Microfor had the consequence that readers who had read the summary would not go back to the full article published in Le Monde; a similar debate took place concerning Google News in several proceedings in various countries. The critics addressed to such reasoning, from a copyright point of view, was that ideas of an article can be taken over and that a summary, if it only takes over ideas and not expressions or (possibly) the structure of the original article should not have been held infringing.

There is yet another aspect to the non-substitutability discussion, which has to do with the method to calculate pricing of access to a database or to a digital scientific periodical: there is a fear that if (1) the number of access is calculated on the basis on the number of times an article has been read and (2) the method does not take into account the cases where the article was downloaded not in order to be read but only in order to be included in a corpus of articles for later analysis (TDM), then this could lead to a loss for the publisher: the articles will have been used, will have added to the quality of the TDM output yet might not have been paid for, if such type of access is not taken into account in the calculation method. These issues however, concern more the agreements between publishers, research institutions and libraries and, even if the fears of the publishers or the libraries are understandable, we do not see that they should be dealt with in a legal provision.

i) Non-applicability to tools designed for data analysis

Publishers and software companies progressively develop tools to facilitate text and data mining. These developments require investments and are of great use to researchers willing to embark on data analysis projects.

A new exception for data analysis should not undermine these developments and investments.

An analogy may be useful here to understand the issue: if an exception to copyright is provided for educational purposes, it is logical (and some Member States have implemented the exception in this manner) to provide that the exception applies to works of literature (novels, fictional books, etc.) but not to school books, i.e. books written for their use by teachers and professors in schools. For such books, an exception allowing their reproduction without having to pay royalties to rightholders undermines the very market of such books. The consequence will logically be that no publishers would be interested in putting such books on the market. The exception could thus make this market disappear.

For data analysis, this means for example that where a database has been put on the market together with software tools allowing TDM within this database, it should be clear that the TDM exception does not apply to these tools and does not allow e.g. to use them without the authorization of the rightholder of said tools. We acknowledge that the distinction between the tools and the data or works may not always

be easy to make and that this requires further analysis. The provision to be drafted for that purpose should not be interpreted as meaning that as soon as a TDM tool is made available by a publisher to mine a database, such database would *ipso facto* stay out of the scope of the exception. A distinction should probably be made between the case where the database cannot be separated from the mining tools and the case where the database could be separated; in the latter case, the publisher should not be able to prevent that such tools be used to do data analysis of the database.

As explained in Part II of our Study, the Japanese exception for data analysis provides for that same principle.

There is maybe another precedent in copyright law which supports this suggestion: the decompilation exception in the Software Directive. Decompilation cannot be excluded by contract, but only if (point b in article 6 hereunder) "*the information necessary to achieve interoperability has not previously been readily available to the persons* (…)".

The article reads as follows:

> *Article 6*
>
> ***Decompilation***
>
> *1. The authorisation of the rightholder shall not be required where reproduction of the code and translation of its form within the meaning of points (a) and (b) of Article 4(1) are indispensable to obtain the information necessary to achieve the interoperability of an independently created computer program with other programs, provided that the following conditions are met:*
>
> *(a) those acts are performed by the licensee or by another person having a right to use a copy of a program, or on their behalf by a person authorised to do so;*
>
> *(b) the information necessary to achieve interoperability has not previously been readily available to the persons referred to in point (a); and*
>
> *(c) those acts are confined to the parts of the original program which are necessary in order to achieve interoperability.*

The underlying idea to point b) above is that if the publisher has provided (under reasonable conditions) the information to the market players having the right to obtain the information, he will not be obliged to accept the reverse-engineering of his software. The analogy cannot be taken too far because the decompilation exception is very specific and relates to competition and interoperability considerations but it may be a useful source of inspiration for policy decisions to be made.

j) Relationship with TPMs

What could be the relationship between TPMs (technical protection measures, as protected by the InfoSoc Directive) and a new TDM exception?

We dealt with this issue above already, when discussed possible legal impediments to data analysis.

Let us recall the solution provided for by article 6.4. of the InfoSoc Directive (even if the provision has up to now largely remained theoretical):

> *"4. Notwithstanding the legal protection provided for in paragraph 1, in the absence of voluntary measures taken by rightholders, including agreements between rightholders and other parties concerned, Member States shall take appropriate measures to ensure that rightholders make available to the beneficiary of an exception or limitation provided for in national law in accordance with Article 5(2)(a), (2)(c), (2)(d), (2)(e),* ***(3)(a),*** *(3)(b) or (3)(e) the means of benefiting from that exception or limitation, to the extent necessary to benefit from that exception or limitation and where that beneficiary has legal access to the protected work or subject-matter concerned.*
>
> *(...)*

> *When this Article is applied in the context of Directives 92/100/EEC and **96/9/EC**, this paragraph shall apply mutatis mutandis." (*emphasis added*).*

The scientific research exception in the InfoSoc Directive is provided by article 5.3.a and thus falls under this mechanism (see article 4, par. 1 above). In accordance with the last paragraph of article 6.4., the Database Directive (Directive 96/9/EC) is also covered and its exception for scientific research thus benefits from the same mechanism.

It means that, in theory, TPMs possibly put in place by rightholders should not be able to prevent the effective benefit of the scientific research exception under the Infosoc Directive and the Database Directive.

We suggest that the same should be true for the specific data analysis exception.

k) Without prejudice to data protection, privacy and confidentiality

As for other exceptions, this exception would only apply without prejudice to other legal rules which may protect certain kinds of data, like personal data (privacy) and confidential data: in such cases, data analysis could not be made by relying on the exception:

- for personal data, because TDM involves "processing" the data in the sense of the personal data protection legislation – and this legislation should continue to apply and restrict the processing of personal data;
- for confidential data, because access thereto necessitates the consent of the person who is in possession of the data: confidential data have not been made available to the public (by definition); if they are protected by copyright, the author has not authorized their divulgation (moral right of divulgation), and exceptions do not apply where works have not been published/made available[335] to the public.

Furthermore, apart from personal data protection and the rules on confidentiality, there may be other rules which may impact and limit the possibilities to do data analysis – see our section on this *supra*.

[335] We use here the words "published" or "made available" in their common sense, not in the legal definition given to them by the Berne Convention or other applicable legislation.

X. CONCLUSIONS AND SUMMARY

In today's world, the amount of available information is growing at an exponential rate, and it becomes more and more difficult to read, on any given topic, even if very specific and narrowly defined, whatever has been published, be it by publishers in subscribed periodicals or databases, in print materials or on the Web. TDM is, according to some, a growing and very promising economic sector. Its applications seem to be full of potentialities, in a whole range of sectors, from forensic investigation, to predictive marketing and scientific research in all kinds of sectors (be they commercial or not). At the same time, in today's world, most information becomes available in a digital format, either from its first creation or because of the growing digitization of existing print archives.

Research is more and more relying on data analysis techniques and, as the quantity of information grows, so does the need to be able to rely on data analysis, because computers and software will more and more have to be called upon to analyze quantities of materials which human beings will, for time and resources reasons, no more have the possibility to read and screen. Data analysis also offers the possibility to uncover new relationships between information and data which science had always been unaware of.

From a legal point of view, data analysis raises various issues, one of the important one being probably privacy and personal data protection. This Study however concentrates only on intellectual property, and particularly copyright and database protection (*sui generis* right).

A few countries in the world have adopted or are in the process of adopting specific copyright provisions to introduce a data analysis exception in their legislation. The purpose of this Study was to describe how data analysis fits within the present legal context in Europe (both in terms of copyright and of database protection), to highlight the issues which may constitute obstacles or difficulties when applying the existing legal texts, and to analyze whether a new exception for data analysis would be useful or necessary. The purpose was not to suggest ready-to-use provisions for a new possible directive but only to give directions and make suggestions.

We first tried to propose a definition of TDM. Regarding terminology, we suggest referring to "data analysis" rather than to "text and data mining" (TDM). "Data analysis" is also the terminology used in the very few legislative texts or drafts presently existing in or out of the European Union. As a definition, we suggest "*the automated processing of digital materials, which may include texts, data, sounds, images or other elements, or a combination of these, in order to uncover new knowledge or insights*".

Information may be available in various ways, from freely available on the Web, to largely available on social networks, only available from publishers to their subscribers, up to (for confidential information) only available after signing non-disclosure agreements. We do not deal with the last category (confidential information). In addition, Open Access models and Creative Commons are also described in our Study to the extent they are relevant to data analysis; and we also mention the EU directives on the re-use of public sector information (PSI) because they facilitate data analysis in this sector of public sector information.

We first examined **which exclusive rights** were involved, from a copyright or database protection perspective:

From a **copyright** perspective, except where the data are in the public domain, because of the way data analysis works, it does, except in very limited cases, involve several acts of **reproduction** of the data. It may be that technical adaptations or translations sometimes also take place, but in a manner in which these acts would also qualify under the broad "reproduction right" of the InfoSoc Directive. Data analysis

does not involve, per se, distribution of the data nor communication to the public – all the more so that one should distinguish the operations involved in the data analysis process itself from what one could then make with the data analysis output (i.e. the report incorporating the results of the analysis): such output is not part of the data analysis process. When the data analysis is made on the basis of data and information held in a database, it will only in some cases but not often involve a reproduction or an adaptation of the database itself and (for the same reasons as explained in the preceding paragraph), it will not involve a communication (or making available) of the database to the public.

From a database protection law standpoint (regarding the *sui generis* rights), when the data analysis is made on the basis of data and information held in a database, it will in most cases involve "**extraction**" of all or substantial parts of the contents of the database but it will normally not amount to "re-utilizing" the same.

Since data analysis does trigger the exercise of several exclusive rights (reproduction for copyright; extraction for the *sui generis* right), we then examined whether **existing exceptions** could apply to data analysis.

We first examined whether the **temporary copy exception** of the InfoSoc Directive (art. 5.1.) could apply:

During each step of a mining process, we saw that potentially numerous copies are made. If any of those copies is permanent, it cannot benefit from the exception.

In only a few rare cases, it could be that copies involved in the steps of a mining process could respect the conditions of the temporary copy. It means that the exception will not provide much relief (or only very rarely) for data analysis activities.

We then examined the **exceptions for scientific research** in both the InfoSoc Directive and in the Database Directive:

In the InfoSoc Directive, if one accepts that the exception does not only apply "to illustrate" scientific research, the exception is useful for data analysis. It does exclude data mining projects which do not qualify as "scientific research", but this does not seem illogical or undesirable. Borderline cases certainly exist between "scientific" and "not scientific" or between "research" and "non research" projects but this does not mean that the distinction is not adequate. The fact that scientific research should be the "sole" purpose may cause concerns as it may exclude projects which should benefit from the exception.

The obligation to mention the source, including the author's name, unless it is impossible, may be rather burdensome and we doubt that it is useful (cf. infra).

The non-commercial purpose criterion may be hard to apply and obviously there are borderline cases. In itself, this is however in our view not an argument to abandon the criterion. When applying the exception to the copyright on a database's selection and arrangement (in the rather unlikely event that data analysis involves their reproduction), the same conclusions may be made than for the exception from the InfoSoc Directive.

When applying the scientific research exception of the *sui generis* right, scientific research must not be the "sole" purpose, but the "sources" must be indicated (no possibility to invoke the fact that this is impossible).

Two specific exceptions of the Database Directive were also examined: (1) the "**normal use of the database**" exception does not give much room for data mining; (2) the **right to extract insubstantial parts** of a database, because the right is only granted to "lawful users", if the subscription agreement to a database prohibits data analysis (or, indirectly, acts which are necessary to perform data analysis), the right cannot be invoked to perform data analysis.

It is worth remembering that, apart from the temporary reproduction exception to copyright, all other exceptions mentioned above are optional. The (not harmonized) manner in which they have been implemented in the Member States may cause difficulties for cross-border projects.

Apart from copyright and database legislation, we listed some **other legislation** which may have an impact on data analysis: data protection legislation must be taken into account, since "mining" data amounts to "processing" them. Contract law and license terms have to be taken into account, except if one suggests that a new data analysis exception cannot be overridden by contract (see hereunder). Legislation in various sectors, on the confidentiality of some data, on data security and on unauthorized access to IT systems may also have an impact on the possibilities to do data analysis. We listed them, without suggesting however that these laws should be amended.

We then analyzed the **initiatives** which the European Commission could take: while encouraging MoUs between stakeholders may prove useful, adopting a simply interpretative document to try to clarify the present legal framework would, in our view, not be sufficient to solve the legal insecurity and remove unjustified obstacles to data analysis.

On the basis of our analysis, **our suggestion** is to have a **new specific data analysis exception** which would be inspired from, and contain partly the same conditions than, the scientific research exceptions, but which would have its own peculiarities. The suggestions are made only for TDM, not for scientific research in general.

The main reasons which, in our view, justify that a specific TDM exception be added to the existing legal framework have been explained in the Study and, in summary, are the following:

- the present situation regarding the research scientific exceptions in the InfoSoc Directive and in Database Directive is not harmonized;
- the exceptions for scientific research can be waived by contract;
- there is a controversy as to whether the exceptions apply to "illustrate scientific research" or "for scientific research as such";
- the copyright exceptions in the InfoSoc Directive are limited "solely for scientific research", which may be seen as excluding projects where, in addition to a scientific research objective, there may be other ancillary objectives;
- the two directives impose mentioning the authors' names and/or the sources, which may not always make sense for data analysis.

On the other hand, the requirement that the exceptions are limited to non-commercial purposes should, in our view, not be changed in a possible new TDM exception.

The specific exception we are suggesting would have its own conditions of application, which would be partly similar to the conditions contained in the other scientific research exceptions and partly different. The specific TDM exception would thus not be subordinate or incorporated in the other scientific research exceptions: it would be a separate exception, to be added to the list of exceptions which already exist.

What should be the **main characteristics and contents** of this new exception?

- the exception would only apply where the purpose is mainly (and not "solely") scientific research;
- it would not just serve "to illustrate scientific research" but would apply more broadly in cases of "scientific research";
- it would only apply where justified by non-commercial purposes;
- mentioning the sources or the names of the authors of the preexisting materials would not be an obligation but be left to the discretion of the researcher(s);
- it would be an exception to the reproduction right (InfoSoc Directive and Database Directive), the adaptation right (both under the Database Direction and as part of the reproduction right in the InfoSoc Directive) and the extraction right (Database Directive);
- it would not apply to tools designed for data analysis (which should remain untouched by the exception);
- it would not apply if the data analysis output substitutes for the pre-existing works or databases and makes the consultation of these pre-existing elements useless;

- it would only benefit users having a lawful access to the data;
- it could not be overridden by contractual terms;
- it would not be optional for Member States but would be mandatory, so as to ensure a level playing field throughout the European Union.

Each of these suggested characteristics is further described and explained in our Study.

We think that the suggested changes would clarify the existing legal framework and would have a positive impact for all stakeholders in the TDM sector, and we hope that they do represent an improvement compared to the present situation and that they provide a balanced solution.

Jean-Paul TRIAILLE

March 2014

BIBLIOGRAPHY

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

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

AIPPI Japanese Group, *Exceptions to Copyright protection and the permitted Uses of Copyright works in the hi-tech and digital sectors* (Question Q216B), 16 May 2011, https://www.aippi.org/download/commitees/216B/GR216Bjapan.pdf

M. BORGHI and S. KARAPAPA, *Copyright and Mass Digitization : a Cross-Jurisdictional Perspective*, Oxford University Press, 2013

J. CLARK, *Text Mining and Scholarly Publishing* , PRC, 2012

J. CLARK*, Text Mining and Scholarly Publishing*, Study Commissioned by the Publishing Research Consortium (PRC), Amsterdam, 2013

Copyright Clearance Center, TDM Pilot Program, *User Guide for CCC's Text and Data Mining Service, updated 1 May, 2013*

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

S. DEPREEUW and J.-B. HUBIN, *Study on the territoriality of the making available right. Localisation of the act of making available to the public and its consequences* in J.-P. TRIAILLE (ed.), Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive"), De Wolf & Partners, European Commission, October 2013, http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf

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

E. DERCLAYE, *The Legal Protection of Databases: A Comparative Analysis*, Edward Elgar, 2008

S. DUSOLLIER, *The Limitations and Exceptions to Copyright and Related Rights for Libraries, Research and Teaching Uses*, in J.-P. TRIAILLE (ed.), Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive"), De Wolf & Partners, European Commission, October 2013, http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf

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

W. FAN, L. WALLACE, S. RICH and Z. ZHANG, *Tapping into the Power of Text Mining,* February 2005

W. J. FRAWLEY, G. PIATETSKY-SHAPIRO and C. J. MATHEUS, *Knowledge Discovery in Databases: An Overview,* 1992

A. GUADAMUZ and D. CABELL *Analysis Of UK/EU Law On Data Mining In Higher Education Institutions* (WHITE PAPER), January 15, 2013

L. GUERNSEY, *Digging For Nuggets Of Wisdom,* October 2013

M. A. HEARST, *What is Text Mining*?, SIMS, UC Berkeley, October 17, 2003

M. A. HEARST, *Untangling Text Data Mining,* May 2012

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

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

JISC , N. KORN, Ch. OPPENHEIM and Ch. DUNCAN*, IPR and licensing issues in derived data*, May 2007

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

JISC, *Use Case Compendium of Derived Geospatial Data*, December 2005

S. JUSOH and H. M. ALFAWAREH, *Techniques, Applications and Challenging Issue in Text Mining,* November 2012

B. LINDNER and T. SHAPIRO, *Copyright in the Information Society*, Edward Elgar, Cheltenham, 2011

B. MICHAUX, *Droit des bases de données*, Kluwer, 2005

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

A. OKERSON, *Text & Data Mining - A Librarian Overview,* August 2013

J.H. REICHMAN and R.L. OKEDIJI, *When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale*, 2012, Minn. L. Rev., I.I.C., vol. 43, 2012

B. F. REILLY, *When Machines do Research, Part 2: Text-Mining and Libraries,* Charleston Advisor, October 2012

P. M. RUST, *The Right to Read Is the Right to Mine*, Open Knowledge Foundation Blog (June 1, 2012), http://bit.ly/O75Rwd

M. SAG, *Orphan works as grist for the data mill*, August 2012

E. SMIT and M. VAN DER GRAAF, *Journal article mining: A research study into Practices, Policies, Plans…..and Promises, May 2011*

E. SMIT and M. VAN DER GRAAF, *Journal article mining: the scholarly publishers'* perspective, Learned Publishing vol. 25 no. 1, January 2012

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

STM, *Text and Data Mining Sample Subscription*, March 2012

STM, *Sample License for Text and Data Mining of subscribed copyright-protected works and materials*, April 2013

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

International Association of Scientific, Technical and Medical Publishers (STM), "*Submission on the Issues Paper, Copyright and the Digital Economy (UK)*, Oxford, 29 November 2012

International Association of Scientific, Technical & Medical Publishers (STM), *Text and Data Mining Sample Subscription*, 15 March 2012, http://www.stm-assoc.org/2012_03_15_Sample_License_Text_Data_Mining.pdf

J.-P. TRIAILLE (ed.), Study On The Application Of Directive 2001/29/EC On Copyright And Related Rights In The Information Society (The "Infosoc Directive"), De Wolf & Partners, Study for the European Commission, October 2013, http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf

J.-P. TRIAILLE, *La question des copies caches et la responsabilité des intermédiaires*, in A. STROWEL and J.-P. TRIAILLE, Google et les nouveaux services en ligne, Larcier, 2008, 257

J.-P. TRIAILLE, *User Generated Content (UGC) – First Part. Description of the present legal situation regarding copyright in the European Union*, *in* Study on the application of Directive 2001/29/EC on copyright and related rights in the Information Society", De Wolf & Partners, Study for the European Commission, October 2013, http://ec.europa.eu/internal_market/copyright/docs/studies/131216_study_en.pdf

M. TRUYENS and P.VAN EECKE, *Legal aspects of text mining*, to be published in *CLSR*, 2014

Universities UK and UK Higher Education International Unit, *European Commission's Stakeholder Dialogue 'Licenses for Europe' and Text and Data Mining*,

http://international.ac.uk/media/2243028/Briefing%20-%20Licenses%20for%20Europe%20and%20Text%20and%20Data%20MiningREVISED.pdf

R. VAN NOORDEN, *Trouble at the text mine, March 2012*

W. WALTER and S. V. VON LEWINSKY, *European Copyright Law : A Commentary*, Oxford University Press, 2010

S. M. WEISS, N. INDURKHYA and T. ZHANG, *Fundamentals of Predictive Text Mining*, Texts in Computer Sciences, Springer, 2010

J. WELLANDER, *Pushing the Frontier on Text Mining: A conversation with Heather Piwowar,* May 2012

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

KM-03-13-426-EN-N

DOI : 10.2780/1475

ISBN : 978-92-79-31976-1