Briefing

Requested by the JURI committee

European Parliament

The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Technical Aspects

KEY FINDINGS

- In the age of 'Big Data' the **value** does not lie in data or text taken in their isolation, but rather in the **extraction** of such value, also by means of predictive text and data mining (TDM) techniques that enable the discovery of **patterns and new relations**.
- **Legal uncertainties** concerning the treatment of TDM practices under EU and national laws **may inhibit the development of TDM in Europe**. Other countries, e.g. **US**, may consider TDM activities as fair use of copyright works. The **UK** has adopted a specific TDM exception, which allows persons having lawful access to undertake text and data analysis for non-commercial research.
- There is consensus around the **potential of TDM in several heterogeneous fields**, also thanks to **affordable technologies**.
- With regard to **business methods**:
    - On consideration of both the variety of TDM practices and their fields of application, the **business models of those that undertake TDM vary**.
    - If one considers the position of **rightholders** – notably publishers – their approaches to TDM differ, with some devising licensing solutions, and others allowing non-commercial TDM without the need for a licence.
- By means of necessary simplification, it appears possible to distinguish **three common steps in TDM processes**:
    - **Access to content**, which may be freely accessible or only accessible through a licence;
    - **Extraction and/or copying of content**, *if required*. In relation to this step, **legal restrictions** – including rights in databases, copyright and related rights – for certain acts of extraction and/or copying might result in the unlawfulness of acts that are preparatory to actual TDM if such acts are carried out without permission from the relevant rightholders.
    - **Mining of text and/or data and knowledge discovery**, which requires the pre-processing of relevant text and data and extraction of structured data, in order to then analyse such output.
- A **mandatory exception** allowing TDM is welcome. Its **scope**, however, should not be unduly narrow and such as to **stifle innovation** coming from different sectors, whether **research organizations or businesses**.

1. THE PURPOSES WHICH CAN BE ACHIEVED BY ENTITIES PRACTISING TDM

In an increasingly data-driven[1] and information-rich socio-economic context, the potential of predictive text and data mining (TDM, sometimes also referred to as text and data analysis) lies in particular in facilitating the processing, recombining, and extraction of further knowledge from **large amounts** of data and text, thus allowing the identification of patterns and associations between seemingly unrelated pieces of information.

Author: Dr Eleonora Rosati, Associate Professor in Intellectual Property Law at the University of Southampton
Research Administrator Responsible: Udo Bux
EN
Policy Department for Citizens' Rights and Constitutional Affairs
European Parliament
PE 604.942

To place things in context, according to an IBM marketing study, 90 percent of the data in the world today has been created in the last two years alone. Every day, 2.5 quintillion bytes of data are created, and it is expected that such growth rate will continue at an even faster pace in the future.[2]

In this sense, the analogy made with the physical universe appears apt: it is expected that by 2020 the digital universe – which consists of data created and copied annually and is doubling in size every two years – will contain nearly as many digital bits as there are stars in the universe. By that time, the digital universe will reach 44 zettabytes, or 44 trillion gigabytes.[3]

In the age of **'Big Data'** data is thus generated in huge amounts, at a very fast pace, and from different sources. The 3 V's of 'Big Data' (volume, velocity, variety) define the classic identity of this phenomenon[4], with further issues – the value and veracity of information – becoming also increasingly central.[5]

Overall, the decreased **cost of data storage**, the deployment of **mobile devices**, **sensors and sensor networks**, as well as the **Internet of Things** are regarded as having contributed to such exponential growth of data.[6]

In addition, also **Artificial Intelligence** and its potential have been acquiring increasing centrality. Although classical **TDM and machine learning** have different utility, it should not be overlooked that both use the **same key algorithms** to discover patterns in data.[7]

The added value and usefulness of TDM

In a context like the one outlined above the value of data does not lie in data or text considered in their isolation, but rather in the **extraction of such value**.[8] This requires that text and data be analysed, to thus enable the discovery of new patterns and relations. Such task would be **virtually impossible to perform manually**, and that is where TDM comes into consideration.

TDM is intended as «a term commonly used to describe the **automated processing** ('machine reading') of large volumes of text and data to uncover new knowledge or insights».[9] Usually – although not always – TDM requires the **copying** of large quantities of material, extracting the relevant data, and recombining it to **identify patterns**.[10]

It should be noted that the research techniques employed for TDM purposes are constantly **evolving** as a result of widespread access to massive, networked computing power and exponentially increasing digital data sets.[11]

The limitations and potential of TDM in Europe

Statistics suggest that **Europe is currently lagging behind** compared to US and Asia (notably China) with regard to number of publications and patents referring to TDM.[12]

According to the EU Commission's Impact Assessment accompany the proposal for a Directive on Copyright in the Digital Single Market[13], there seems to be general agreement among stakeholders that TDM «is still a **nascent tool, in particular in the non-business sector**, i.e. for research carried out by organisations such as universities or research institutes».[14]

EU-based researchers indicate a number of **factors that inhibit the development of TDM** in Europe, including:

- **legal uncertainties** concerning the treatment of TDM activities under EU and national copyright laws;
- lack of awareness and skills; and
- infrastructural challenges.[15]

Nonetheless, there appears to be consensus around the **potential of TDM**.[16] Not only does its application span **several heterogeneous fields**, but it also appears that «[b]ecause TDM research technology is not prohibitively expensive, it is **readily available** to lone entrepreneurs, individual post-graduate students, start-ups and small firms».[17]

The beneficiaries

At the moment it is difficult to determine the actual amount and identity of those undertaking TDM activities. While it appears that employment of TDM in certain fields is limited, there are research areas in which TDM plays a significant role.

In the field of **computational linguistics** (or human language technology, natural language process), TDM accounts for about 25-30% of all research projects.[18] A range of case studies suggests that the use of TDM has increased efficiencies and the speed of **biomedical discovery**. For instance, research utilizing text mining has facilitated the «creation of hypotheses regarding the roles of four genes never previously

characterised as involved in craniofacial development a significant breakthrough with far reaching implications».[19]

The applications of TDM techniques are **not limited to the field of research**. TDM is employed in fact in **a variety of heterogeneous business fields**, that may range from forestry methods to banking, from marketing to criminology, from anthropology to fashion, and so on.

In this sense, for instance, data mining may be used to manage and optimize the use of **natural resources**.[20] It may assist **banks** in identifying and quantifying the accuracy, timeliness, and forward-looking character of their credit-risk-assessment systems, as well as traditional analyses of industries and sectors.[21] TDM may serve in **marketing** to visually track information spread and influence paths across media segments in real-time, along with improving content relevance for target audiences.[22] As regards its employment in the field of **criminology**, it is believed that TDM can help exploring and detecting crimes and their relationships with criminals, so to assist police forces.[23] By mining the ever-growing content made available through social networks, it is believed that the information thus retrieved may assist **anthropologists** in their studies of cultural phenomena.

Recent research conducted at **Cornell University** with the aid of TDM techniques has for instance allowed a group of researchers to mine 100 million photographs made available on **Instagram** and devise patterns on how **clothing styles** vary around the world, and tackle the frequency of use of certain garments and colours. By training a **machine-learning algorithm**, the researchers were able to identify a set of visual themes and study how these would vary by time and place, and also identify the preference for certain colours.[24]

A significant aspect to note is also that the **same technology may perform TDM for different purposes**: for instance, **IBM Watson Explorer** has been used – among other things – to:

- Improve productivity in the workplace;
- Increase efficiency in public health management, e.g. in Italy[25];
- Improve diagnostic information, as has been the case for oncologists working at the Memorial Sloan-Kettering Cancer Centre in New York[26];
- Allow for tailored drug recommendations, by reducing the time needed to create a report based on the single patient's own gene alterations related to a tumour from a week to a few minutes;
- Prevent the commission of crimes, including cybercrimes and hacks;
- Create new culinary recipes[27]; and
- Predict (correctly) who would win Italy's best-known singing competition (Sanremo 2018).[28]

Most common business models

On consideration of the variety of TDM methods, their fields of application, and the fact that in certain areas use of TDM is arguably in its infancy, **the business models of those that employ TDM techniques vary**.

If one considers instead the position of **rightholders** – notably publishers –, their business models have been traditionally rooted within control of access to and exploitation and re-use of content.

In this sense, the principal question arisen in relation to TDM is whether such activities could be **authorized** by relevant rightholders or could be, instead, subject to a **copyright exception** (as is the case, for instance, in the UK with section 29A and Schedule 2(2)1D of the Copyright Designs and Patents Act 1988[29], and as the Irish Copyright Review Committee proposed in the context of a review of Irish copyright law in 2012[30]), or even be **outside the scope of copyright protection *tout court*** (according to the idea for which «The right to read is the right to mine».[31]).

Certain types of publishers maintain a sceptical outlook on allowing TDM activities, including by means of a general licensing policy. In this sense, newspaper publishers have held the view that **licensing should take place on a case-by-case basis** even for non-commercial research to prevent massive abuse or loss of their archives and the destruction of their business model.[32]

With regard to **academic publishers**, the trend that is being observed is the attempt –by some of them – to offer the possibility of undertaking TDM activities as part of their **licensing models**. For instance, **Elsevier** makes available all its journals and book chapters in XML, this being a normalized format that allows the undertaking of TDM activities. Users can access and register an account on a developers' portal and create an online interface (API key) that serves as a first step for performing TDM activities.[33]

Certain academic publishers like **Oxford University Press** and **Springer** allow the undertaking of TDM activities for **non-commercial use**, **without the need for a licence**. According to the relevant policies[34], permission is not required for non-commercial text mining. It appears that this would be the case even if the relevant acts of copying that are functional to the performance of TDM take place outside territories, e.g. the UK, where specific copyright exceptions allowing this type of TDM are specifically in place.[35]

In relation to TDM for **commercial use**, a number of publishers (including Oxford University Press and Springer Nature, as well as Wiley, BMJ, the Royal Society of Chemistry, Taylor & Francis, SAGE, Cambridge University Press, American Diabetes Association, American Society for Nutrition, and Future Medicine) participate in the Copyright Clearance Center's **RightFind™ XML for Mining solution**. This allows perspective users of content for commercial TDM purposes to obtain XML-formatted content on demand from publications they subscribe to and discover unsubscribed published material.[36]

2. THE PROCEDURE OF TDM AND THE STATE OF THE ART

TDM activities can take place through **different procedures and with different goals**, the **only** common element being that of **analysing and extracting associations between concepts to identify new patterns and relations**.

With regard to text mining, an example of the different types of mining procedures and end-goals available is provided by the various services made available by **NaCTeM**, the national UK centre for text mining and world' first publicly funded centre for the development of tools and services to support the UK academic community.[37]

Different steps in the process of TDM

The variety of TDM techniques, practices and end-goals makes it virtually impossible to provide a general and exhaustive illustration of how TDM works.

By means of a necessary simplification, it appears however possible to distinguish **three common – yet not all necessary – steps** in TDM processes:

1. Access to content (**Step 1**);
2. Extraction and/or copying of content (**Step 2**);
3. Mining of text and/or data and knowledge discovery (**Step 3**).

As it will be explained further below, legal issues might arise in TDM processes that include, in particular, Step 2.

The following sections explain the technical aspects relating to each and every step, as well as outlining the possible legal issues that might arise in connection with each of them.

• Step 1 – Access to content

The first and propaedeutic step to TDM activities is access to the content – whether text or data – in relation to which TDM is to be performed.

The primary distinction to be made is between content that is **freely accessible** and content that is not, and in relation to which access permission, i.e. a **licence**, may be required.

In relation to the former, **freedom of access does not necessarily entail that the content (text and data) is also free of legal restrictions**. As it will be explained further below sub Step 2, in fact, different legal rights might be vesting on the content that needs to be extracted and/or copied.

In relation to the latter, an issue might be also that of **identifying the subjects from whom permission is to be sought**, i.e. the relevant rightholders.

In the particular context of **out-of-commerce works**, it is worth recalling that recently the **Court of Justice of the European Union (CJEU)** has stressed that the undertaking of relevant restricted acts is subject to the **actual and prior consent of relevant rightholders (authors)**: without guarantees that these are actually informed as regards the envisaged use of their works, other subjects (e.g. collective rights management organizations) are not in a position to adopt any position whatsoever as to such use.[38]

Problems might also arise in relation to **orphan works**, these being works and other subject-matter that are protected by copyright or related rights and for which no rightholder has been identified or for which the rightholder, even if identified, has not been located.[39]

In the context of a rights clearance simulation study, for instance, the **British Library** estimated that of the total number of potentially in-copyright works in its collections over 40% might be orphan works.[40] Before the adoption of legislative initiatives at the EU and UK levels, the **BBC** stated that over one million hours of television programming from its archives were not used due to the impossibility, or the disproportionate cost, of tracing rightholders, as well as the risk of subsequent legal actions.[41]

In the course of an empirical study devoted to investigating the legal challenges that UK **cultural heritage institutions** have encountered in undertaking the digitization of works in their own collections and archives, it was found – at least before the adoption of the EU Directive on certain permitted uses of orphan works (Directive 2012/28/EU[42], which introduced new mandatory exceptions on orphan works) and the 2013 UK Enterprise and Regulatory Reform Act (which introduced a licensing scheme for the use of orphan works) – that UK cultural institutions had developed and adopted **risk management strategies** to overcome the obstacles posed by the impossibility of identifying or locating relevant rightholders. While some institutions did not consider the impossibility of identifying or locating relevant rightholders as an absolute bar to the undertaking of digitization projects and making digitized content available to the public (e.g. via the concerned institution's official website, as the **Wellcome Trust** did), other institutions, including the **National Library of Scotland** and the **Imperial War Museum**, avoided engaging in digitization activities altogether lacking complete rights clearance *ex ante*, while others (e.g. the **British Museum**) used these works only internally.[43]

If a licence is required and is successfully secured, its resulting scope determines the **types of activities that the licensee is entitled to undertake** in relation to the content to which access has been secured. It is worth recalling that **exceeding the scope of the licence secured might expose the licensee to liability** for infringing acts.

As mentioned above, **some publishers include the possibility of undertaking TDM activities within the scope of the licences available, but that is not always the case**. In particular, if acts of extraction and/or copying of content are needed to undertake TDM, then further issues should be considered by the licensee who is not also explicitly allowed to perform TDM on the licensed content.

- ## Step 2 – Extraction and/or copying of content

Lawful access to content – whether because such content is freely accessible or access has been obtained through a licence – **does not necessarily entitle one to undertake TDM** in respect of such content (text or data).

This is because **to undertake TDM it may be necessary to undertake certain propaedeutic activities**, including extracting and/or copying the content, **for which specific authorization may be required.**

In any case, it is necessary to stress at the outset that:

- While what is stated above holds true for some TDM techniques, **not all TDM practices require necessarily the extraction and/or copying of content**. This may be because, for instance, the TDM technique employed does not require undertaking such activities at the outset.

- In addition, **not all acts of copying are necessarily subject to the control of the relevant rightholder**. If the content being temporarily copied satisfies, e.g., the conditions under the only mandatory exemption under the InfoSoc Directive (Directive 2001/29/EC[44]), this being Article 5(1), then such acts of copying would not require permission from the relevant rightholder. As the **UK Supreme Court** has clarified, «[m]erely viewing or reading it is not an infringement.»[45] This is true as long as, among other things, the person making the temporary copies has **lawful access** to the content being copied. As the **CJEU** has held on a number of occasions, «a use should be considered lawful where it is authorised by the right holder or where it is not restricted by the applicable legislation».[46] In this sense, for instance, it is arguable that activities like **web mining**, which involve the application of data mining techniques to extract knowledge from web data (including web documents, hyperlinks between documents, usage logs of web sites, etc)[47], might not always require authorization from relevant copyright holders.

If the TDM technique employed involves the making of acts of extraction and/or copying in scenarios other than those discussed above, then **legal restrictions** might be in place.

If the content extracted and/or copied is included in a **database**, then both copyright and the sui generis (database) right might come into consideration, as well as other aspects in the event that neither vests in the database considered.

It should be recalled that in its case law on the Database Directive (Directive 96/9/EC[48]) the CJEU has **defined the concept of 'database' broadly**, in the sense of:

- Referring to databases 'in any form' and irrespective of whether they are in electronic or non-electronic format; and

- Applying to literary, artistic, musical or other collections of works or collections of other material such as texts, sound, images, numbers, facts, and data.[49]

As mentioned, both copyright and the sui generis right (these being two rights that are independent of each other and may subsist together at the same time on the same database) may come into consideration in the case of content included in a database.

With regard to **copyright**, the **author of a database** is entitled to prevent a number of acts, including the reproduction – whether **temporary or permanent** – by any means and in any form, in whole or in part of the expression of the database which is protectable by copyright, i.e. expression that is sufficiently original. The only mandatory limitation to the rights of the copyright holder relates to the performance by the lawful user of a database or of a copy thereof of any acts that are necessary for the purposes of accessing the content of the databases and its normal use.

With regard to the **sui generis (database) right**, the **maker of a database** who has made qualitatively and/or quantitatively a substantial investment in either the obtaining, verification or presentation of the contents is entitled 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.

Restrictions may also subsist in relation to **databases that are protected by neither copyright nor the sui generis (database) right**. As the CJEU has clarified[50], in fact, the holder of a database of this kind is not subject to the limitations to copyright and the sui generis right set in the Database Directive. Hence, he/she is **free to determine by contract and in compliance with the applicable national law the conditions of use of its database**.

In addition to legal issues and rights vesting in databases, subjects with access to content that they intend to use to TDM purposes will also need to consider issues of **copyright and related rights**, potentially including – should the proposal contained in Article 11 of the draft Directive on Copyright in the Digital Single Market be adopted – a new press publishers' right.

With regard to **copyright**, if the reproduction at issue does not satisfy the conditions for the application of the exemption within Article 5(1) of the InfoSoc Directive, then consideration should be given to the fact that the CJEU in its case law has defined the **concept of reproduction** – including reproduction 'in part' of a work – **broadly**. The type of **assessment** required is in any case qualitative, rather than quantitative: there is an act of reproduction anytime what is being reproduced is **sufficiently original**, in the sense of being its author's own intellectual creation.[51]

Also **related (neighbouring) rights** might come into consideration in relation to perspective acts of copying finalized to the undertaking of TDM activities. Should the **proposed press publishers' rights** be ultimately adopted, acts of reproduction in respect of press publications might also require authorization of press publishers. From the progress of discussion at this stage, this could be so also irrespective of whether the reproduction at issue concerns content that is sufficiently original in a copyright sense.

Finally, it should be noted that not only might intellectual property rights limit the activities underlying Step 2, but also other areas of the law might be relevant at this stage. In this sense, the application of **data protection and privacy laws** to the realm of text and data extraction should be considered, including relevant provisions on the processing of personal data for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in the **General Data Protection Regulation** (Regulation (EU) 2016/679[52]). Another area that might be relevant is **contract law**, especially in relation to contractual restrictions and – where applicable – contractual restrictions of TDM (in this respect, it is worth highlighting that the UK TDM exception expressly prevents **contractual override**).

- ### Step 3 – Mining of text and/or data and knowledge discovery*

Following the preliminary steps illustrated above, the actual **mining** activity takes place. This step is also **propaedeutic** to the realization of the goal underlying predictive TDM, which is not just mere extraction of information, but rather **knowledge discovery**.

In addition to the steps discussed above, which consist of identifying the content to use and securing access to it, in most cases stages in text and data mining processes include[53]:

A. Pre-processing of relevant text and data (**Stage A**);

B. Extraction of structured data (**Stage B**).

The starting point of TDM activities is the **pre-processing of the relevant text and data (Stage A)**. It might happen that text is in an unstructured form and input data is 'noisy' and ridden with errors. Text and data **clean-up** is thus necessary. **Cleaning and parsing** are required because text is created and stored in such a way that, while understandable to humans, it may not be in a form amenable to having the encoded meaning easily extracted by a computer.[54] In general terms this phase consists of:

- The removal of any unnecessary or unwanted information, e.g. advertising;

* In relation to this part the Author is grateful for the feedback and comments provided by Dr Jonathon Hare (University of Southampton – School of Electronics & Computer Science). Errors and omissions are solely of the Author.

- Dealing with tables, figures and formulas.

The **normalization** of the text and data, e.g. the conversion into XML format, may be required but that may not always be the case.

The subsequent process of **extraction (Stage B)** requires:

- '**Tokenization**' to break the document into its constituent terms, while possibly also eliminating certain characters, e.g. punctuation. For some languages this operation is straightforward (English, French, etc), but for others (Chinese, etc) it is non-trivial;

- The **identification of synonyms** through linguistic resources (e.g. processes such as 'lemmatization'), or rule-driven approaches (e.g. 'stemming'), or by learned statistical approaches ('embedding');

- The following step involves **text transformation (attribute generation)**, which involves the representation – usually as bag of words or vector space – of the text document through the words contained and their frequency of use. While this step is usually present in many classical models, more recent approaches make **use of the raw text as their input** (the classical 'bags of words' throw away the ordering of terms and, as such, may lose context);

- Identifying **equivalence classes**. Choosing 'weights' to make some classes more important than others may help do this.

In the particular context of **extraction**, many modern approaches are almost exclusively **statistical** in nature and do not necessarily require any linguistic approaches such as natural language processing (NLP), i.e. computational techniques for the analysis and synthesis of natural language and speech, to extract concepts from the examination of words, phrases, and syntax, or structure, of text.

In the context of **predictive data mining** in which 'supervised' machine learning approaches are employed, Stage B is followed by the identification (and verification) of **patterns and events extraction**. TDM practices can thus allow the discovery of not just types and concepts but also relationships among them and patterns. This way, TDM might be also used as a tool to predict future or otherwise unknown events (**predictive analytics**).

- ### Summary of the various steps

In light of the discussion above, the various steps involved in TDM processes that may be relevant from a legal perspective may be simplified and summarized as proposed in the charts *sub* **Annex I.**

3. THE BORDER LINE TO INTELLECTUAL PROPERTY RIGHTS

As discussed above, intellectual property rights may affect and hinder TDM activities. Although in certain fields TDM is arguably in its infancy, there appears to be consensus around its potential and increasingly central relevance.

Among the reasons provided to justify the relatively limited use of TDM in Europe, there is **uncertainty regarding the relevant legal regime** under EU copyright and database rules. In this sense, a comparison with jurisdictions that display a more frequent and advanced use of TDM has been at times made, in order to highlight the differences in the legal regulation of TDM.

The US approach

The **US copyright regime** is considered more favourable to TDM practices than what appears to be the case under European laws, also because of the inherent flexibility of the **fair use doctrine** under §107 of the US Copyright Act as opposed to a EU-style closed system of exceptions and limitations.

Under the US fair use doctrine – generally speaking – it is not required to identify a specific exception or limitation that could be invoked as a defence against a finding of *prima facie* infringement. What is required is, instead, the consideration of a limited number of factors in order to come to a determination of whether a certain use made of a copyright work is fair. Defining the scope of fair use under US copyright law has been arguably a matter of judicial application and case law development.

The fair use assessment requires to consider - among other things - whether the use made of a work «adds value to the original - if the quoted matter is used as raw material, transformed in the creation of new information, new aesthetics, new insights and understandings - this is the very type of activity that the fair use doctrine intends to protect for the enrichment of society».[55]

In the context of the present discussion, it may be worth recalling that, in the longstanding litigation over the **Google Books Library Project**, it was considered relevant for a finding of fair use also the fact that the search engine «makes possible new forms of research, known as "text mining" and "data mining."» by using the Google Library Project corpus «to furnish statistical information to Internet users about the frequency of word and phrase usage over centuries».[56]

There is case law in the US that suggests that **acts of incidental or intermediate copying** which do not ultimately result in the external re-use of protectable (expressive) parts of a copyright work should not be considered infringing, i.e. such as to **supersede the objects or purposes of the original creation**.[57] It has been argued that **TDM activities should be considered in this perspective**, i.e. as acts of incidental/intermediate copying that are finalized to accessing (and using) the unprotected parts of a copyright work.[58]

The UK approach

If one considers the regulation of TDM in Europe, the UK has been **the first EU Member State** (more recently, also France, Estonia and Germany have legislated in this area) to introduce a specific copyright defence allowing **text and data analysis for non-commercial research**. As of today, the provision contained in section 29A of the Copyright, Designs and Patents Act 1988 has not yet received judicial application.

However, it should be noted that, while the defence is limited to the copying of a work finalized to the undertaking of a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, the **beneficiaries of the exception are any persons who have lawful access to the work in question**.

At the time of introducing such exception into UK law (2014) within the framework offered by **Article 5(3)(a) of the InfoSoc Directive**, UK Government believed that:

- The resulting defence should be framed within a **scientific research purpose**, and be allowed insofar as this would remain of a **non-commercial character**. This would be necessary to comply with the wording of **Article 5(3)(a) of the InfoSoc Directive**;

- **No restrictions should be imposed on the beneficiaries** of the resulting exception, as long the work used to make a copy for text and data analysis research is one to which the relevant person has lawful access.

Although UK Government acknowledged «that some publishers take an active role in developing text and data analytic technologies, and that some offer contracts that support the use of these technologies», «under current conditions, research projects may in some cases require **specific permissions from a large number of publishers** in order to proceed», and that this «is in some cases an **insurmountable obstacle**, preventing a potentially significant quantity of research from taking place at all».[59]

In light of the evidence collected during the public consultation phase, UK Government estimated that:

> «a permitted act **would benefit researchers by £124m each year**, and given the strength of the UK research base and the growing importance of data sharing/re-use this has a potential to be much higher if exploited effectively. The copying involved in text and data analytics is a necessary part of a technological process, and is **unlikely to substitute for the work in question** (such as a journal article). It is therefore **unlikely that permitting mining for research will of itself negatively affect the market for or value of copyright works**. Indeed, it may be that removing restrictions from analytic technologies would **increase the value** of articles to researchers».[60]

The EU approach?

In 2016 the EU Commission issued a proposal for a Directive on Copyright in the Digital Single Market which, among other things, includes a new mandatory exception allowing **research organizations** to make reproductions and extractions in order to carry out TDM of works or other subject-matter to which they have lawful access for the purposes of scientific research.

Although including **commercial and non-commercial uses alike** (thus differing from the UK TDM exception), the **catalogue of beneficiaries would be narrower** than what is the case under the UK exception. By 'research organizations', it is in fact intended universities, research institutes, non-profit or public interest research-intensive organizations.

In principle the draft directive does not exclude applicability of the text and data mining exception to **public-private partnerships**, but rules out that this could be possible when a commercial undertaking has a **decisive influence and control** over the research organization in question.

Following the release of the proposal, **different approaches** to a EU TDM exception have been proposed. Among other things, in their Opinions for the Committee on Legal Affairs (JURI) of the European Parliament:

- The **Committee on the Internal Market and Consumer Protection (IMCO)** has advised to clarify that no authorization would be required in cases where TDM is carried out in relation to mere facts or data which are not protected by copyright, because «**The right to read is in effect the same as the right to mine**». In addition, copies made for TDM should not be **stored** for a longer time than what is necessary for TDM purposes, and beneficiaries should include research organizations and **cultural heritage institutions** alike.[61]

- The **Committee on Industry, Research and Energy (ITRE)** has suggested adding language to the text of the TDM exception in the sense of allowing rightholders to implement **measures** where there is risk that the security and integrity of the system or databases where the works or other subject-matter are hosted would be jeopardized. At the same time, such measures should **not prevent or exclude the ability to develop TDM tools** different from those offered by the rightholder as long as the security and integrity of the networks and databases is protected. **No fair compensation** requirement would be needed. In addition, the exceptions should be also available to **start-up companies** that satisfy certain requirements of size, turnover, and date of establishment.[62]

- The **Committee on Culture and Education (CULT)** has suggested strengthening the language of Article 3, in the sense of prohibiting rightholders from **hindering** research organizations from availing themselves of the TDM exception, and envisaging the possibility of a **fair compensation** requirement to be left to the discretion of individual Member States.[63]

At the time of writing, the Opinion of the Committee on Legal Affairs (JURI) is awaited.

CONCLUSION

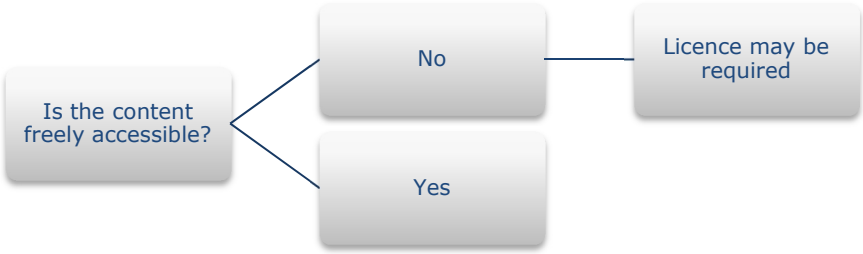
Although less developed than in other jurisdictions, TDM techniques display a **clear potential** also for European researchers and businesses.

The **legal issues** outlined above may, however, **hinder these practices**, and make rights clearance burdensome in relation to processes which may include the extraction and copying of content as an **incidental and intermediate phase** of an activity which, as a whole, intends to extract ideas and information from such content and use them for non-expressive purposes.

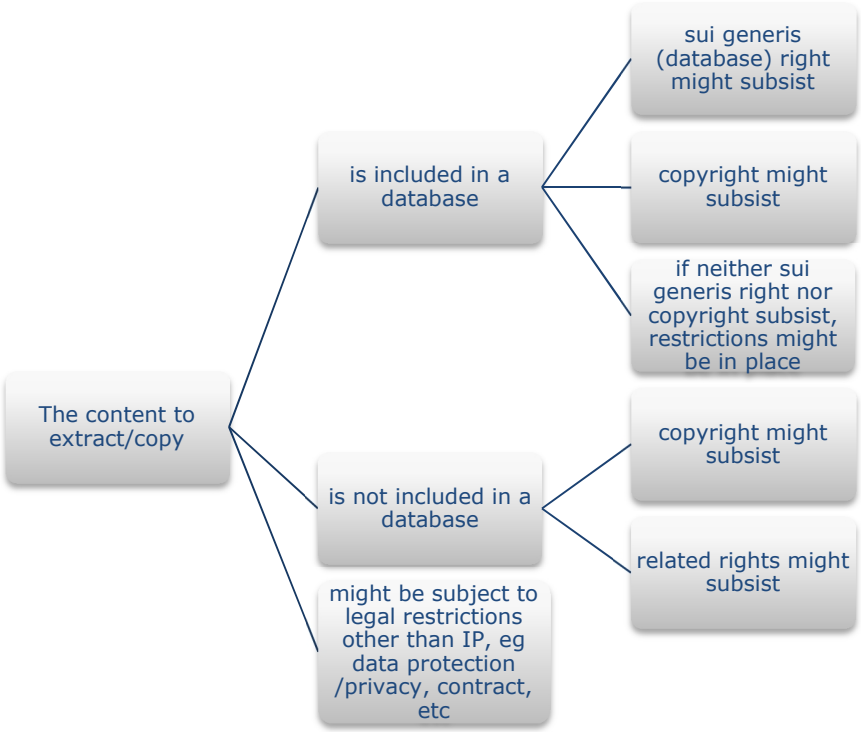
A **mandatory exception** allowing TDM is a welcome proposal. Its **scope**, however, should not be unduly narrow and such as to stifle innovation coming from different sectors, **whether research organizations or businesses**. In this sense, the EU legislature should carefully consider who the **beneficiaries** of the resulting exception should be, as well as the **uses** allowed of works or other subject-matter for TDM purposes.

ANNEX I

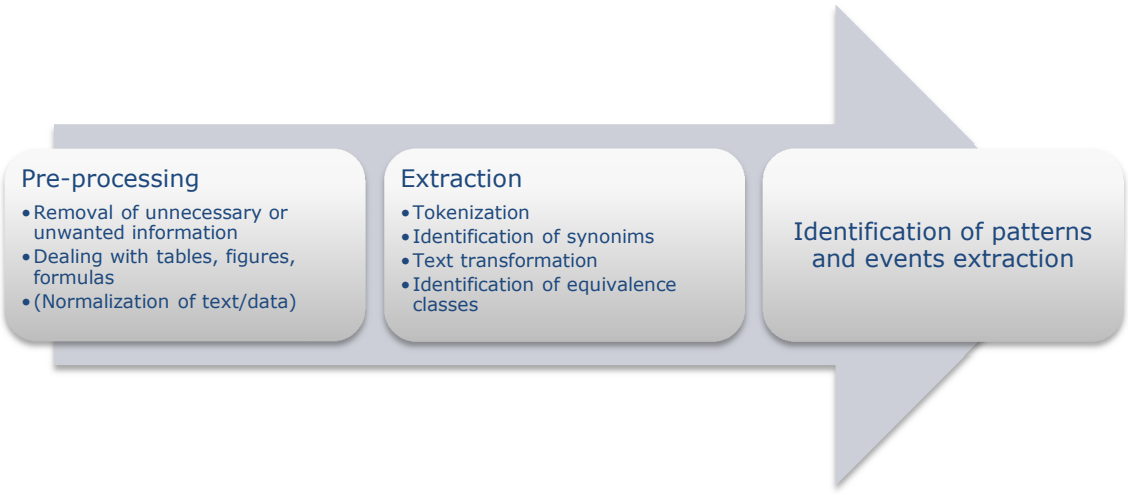
Step 1 – Access to content



Step 2 – Extraction and/or copying of content



Step 3 - Mining of text and/or data and knowledge discovery



[1] OECD (2013), «Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by "Big Data"», OECD Digital Economy Papers, No. 222, OECD Publishing, Paris. http: //dx. doi. org/ 10.1787/5k47zw3fcp43-en,%204.
[2] IBM Marketing Cloud (2017), «10 Key Marketing Trends for 2017 and Ideas for Exceeding Customer Expectations».
[3] IDC (2014), «The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things».
[4] Eaton *et al* (2012), «Understanding Big Data – Analytics for Enterprise Class Hadoop and Streaming Data» (McGraw Hill), 5.
[5] Demchenko *et al* (2013), «Addressing Big Data Issues in Scientific Data Infrastructure», 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, 2013, 48-55.
[6] Filippov (2014), «Mapping Text and Data Mining in Academic and Research Communities in Europe» (The Lisbon Council), 3.
[7] Brooks *et al* (2017), «Artificial Intelligence vs. Machine Learning vs. Data Mining 101 – What's the Big Difference?» (Guavus Blog).
[8] IDC (2014), «The Digital Universe», *cit*.
[9] EU Commission (2016), «Commission Staff Working Document – Impact Assessment on the Modernisation of EU Copyright Rules Accompanying the Document Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council laying down Rules on the Exercise of Copyright and Related Rights Applicable to Certain Online Transmissions of Broadcasting Organisations and Retransmissions of Television and Radio Programmes», SWD(2016) 301 final, Part 1/3, §4.3.1.
[10] UK Intellectual Property Office (2011), «Supporting Document T Text Mining and Data Analytics in Call for Evidence Responses».
[11] Hargreaves *et al* (2014), «Standardisation in the Area of Innovation and Technological Development, notably in the Field of Text and Data Mining – Report from the Expert Group», doi:10.2777/71122, 10.
[12] Filippov (2014), «Mapping Text and Data Mining», *cit*, 9 and 13.
[13] Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM(2016)593.
[14] EU Commission (2016), «Commission Staff Working Document», *cit*, §4.3.1.
[15] *Ibid*.
[16] *Ibid*.
[17] Hargreaves *et al* (2014), «Standardisation in the Area of Innovation and Technological Development», *cit*, 10.
[18] Filippov (2014), « Mapping the Use of Text and Data Mining», *cit*, 28.
[19] Leach *et al* (2009), « Biomedical Discovery Acceleration, with Applications to Craniofacial Development », PloS Computational Biology 5(3): e1000215. doi:10.1371/journal. pcbi.1000215, as cited in UK Intellectual Property Office (2011), «Supporting Document T Text Mining and Data Analytics», *cit*, 2.
[20] See, e.g., http://www.treemetrics.com/ (accessed 14 February 2018).
[21] McKinsey & Company (2013), «Ratings Revisited: Textual Analysis for Better Risk Management».
[22] See, e.g., https://www.ubermetrics-technologies.com/influencers-content-marketing/ (accessed 14 February 2018).
[23] RezaKeyvanpour *et al* (2011), «Detecting and Investigating Crime by means of Data Mining: A General Crime Matching Framework», doi.org/10.1016/j.procs.2010.12.143.
[24] Matzen et al (2017), «StreetStyle: Exploring World-Wide Clothing Styles from Millions of Photos» arXiv:1706.01869 [cs.CV].
[25] See https://tinyurl.com/yadrpf6n (accessed 14 February 2018).
[26] See http://www.wired.co.uk/article/ibm-watson-health-cancer-kyu-rhee (accessed 14 February 2018).
[27] See http://www.wired.co.uk/article/ibm-watson-artificial-intelligence (accessed 14 February 2018).
[28] See https://tinyurl.com/ycgwms63 (accessed 14 February 2018).
[29] See http://www.legislation.gov.uk/uksi/2014/1372/regulation/3/made (accessed 14 February 2018).
[30] Copyright Review Committee (2012), «Copyright and Innovation – A Consultation Paper», §9.8.
[31] Murray-Rust (2012), «The Right to Read is the Right to Mine» (Open Knowledge International Blog).
[32] ENPA written response to the DG Research Expert Group on Standardisation (2014), as reported in Hargreaves *et al* (2014), «Standardisation in the Area of Innovation and Technological Development», *cit*, 18.
[33] See https://www.elsevier.com/authors-update/story/access-to-research/demystifying-text-and-data-mining and https://dev.elsevier.com/ (both accessed 14 February 2018).
[34] See https://tinyurl.com/y85lvysp and https://tinyurl.com/y9s9gv4m (accessed 14 February 2018).
[35] With regard to the UK exception, according to Jisc «Researchers who are based abroad [i.e. outside the UK], and not affiliated to a UK institution would need to refer to the copyright law of their own jurisdiction for the equivalent exception. Where the researcher is based abroad, is registered with, and has lawful access to the licensed content via the UK institution, then making copies for computational analysis under the exception should only be done in the UK by UK based colleagues» (Jisc (2016), «The Text and Data Mining Copyright Exception: Benefits and Implications for UK Higher Education»).
[36] See http://www.copyright.com/business/xmlformining-2/ (accessed 14 February 2018).
[37] See http://www.nactem.ac.uk/services.php (accessed 14 February 2018).
[38] CJEU, *Marc Soulier and Sara Doke v Premier ministre and Ministre de la Culture et de la Communication*, C-301/15, EU:C:2016:878
[39] See Article 2(1) of Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works, OJ L OJ L 299, 5-12; see United States Copyright Office, «Report on Orphan Works. A Report of the Register of Copyrights» (2006), 1.

[40] Stratton (2011), «Seeking New Landscapes. A Rights Clearance Study in the Context of Mass Digitisation of 140 Books Published between 1870 and 2010», 5
[41] Kroes (2011), «Addressing the Orphan Works Challenge», SPEECH/11/163.
[42] Directive 2012/28/EU of the European Parliament and of the Council of 25 October 2012 on certain permitted uses of orphan works Text with EEA relevance, OJ L 299, 5–12.
[43] Rosati (2013), «Copyright Issues facing Early Stages of Digitization Projects» (University of Cambridge), 4.
[44] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society, OJ, L 167, 10-19.
[45] UKSC, *Public Relations Consultants Association Ltd v The Newspaper Licensing Agency Ltd*, [2013] UKSC 18, paragraph 1 (Lord Sumption).
[46] CJEU, *Stichting Brein v Jack Frederik Wullems*, C-527/15, EU:C:2017:300, paragraph 65, referring CJEU, *Football Association Premier League and Others*, C-403/08 and C-429/08, EU:C:2011:631, paragraph 168, and CJEU, *Infopaq International*, C-302/10, EU:C:2012:16, paragraph 42.
[47] Srivastava et al (2005), «Web Mining – Concepts, Applications and Research Directions» (Springer), 399.
[48] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 20–28.
[49] See, e.g., CJEU, *Freistaat Bayern v Verlag Esterbauer GmbH*, C-490/14, EU:C:2015:735, paragraphs 13-14, referring to CJEU, *Fixtures Marketing*, C-444/02, EU:C:2004:697, paragraph 23.
[50] CJEU, *Ryanair Ltd v PR Aviation BV*, C-30/14, EU:C:2015:10.
[51] CJEU, *Infopaq International A/S v Danske Dagblades Forening*, C-5/08, EU:C:2009:465.
[52] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ L 119, 1–88. For a discussion of the interplay between TDM and data protection and privacy laws before the adoption of Regulation (EU) 2016/679, see Hargreaves *et al* (2014), «Standardisation in the Area of Innovation and Technological Development», *cit*, 61-64.
[53] See https://tinyurl.com/yd5t98j4 (accessed 14 February 2018).
[54] See https://guides.library.duke.edu/c.php?g=289707&p=1930855 (accessed 14 February 2018).
[55] Leval (1990), «Toward a Fair Use Standard» 1989-1990 103 Harv L Rev 1105, 1111.
[56] *Authors Guild v Google, Inc*, No. 13-4829 (2d Cir. 2015), affirming *Authors Guild v Google, Inc*, 954 F.Supp.2d 282 (2013).
[57] See Sag and Schultz (2013), «Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v Hathitrust», 4, referring to, e.g., *A.V. ex rel. Vanderhye v. iParadigms, LLC*, 562 F.3d 630, 645 (4th Cir. 2009); *Perfect 10, Inc. v. Amazon.com, Inc*., 508 F.3d 1146, 1168 (9th Cir. 2007); *Sony Computer Entm't, Inc. v. Connectix Corp*., 203 F.3d 596, 609 (9th Cir. 2000); *Sega Enters. Ltd. v. Accolade, Inc*., 977 F.2d 1510, 1527-28 (9th Cir. 1992)
[58] *Ibid*, 24-31.
[59] HM Government (2012), «Modernising copyright: a modern, robust and flexible framework. Government response to consultation on copyright exceptions and clarifying copyright law», 37.
[60] *Ibid*.
[61] Opinion of the Committee on the Internal Market and Consumer Protection for the Committee on Legal Affairs on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market 2016/0280(COD) (Rapporteur: Catherine Stihler), in particular Amendments 4, 5, 8, 40, 42.
[62] Opinion of the Committee on Industry, Research and Energy for the Committee on Legal Affairs on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market 2016/0280(COD) (Rapporteur: Zdzisław Krasnodębski), in particular Amendments 5, 6, 7, 35.
[63] Opinion of the Committee on Culture and Education for the Committee on Legal Affairs on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market 2016/0280(COD) (Rapporteur: Marc Joulaud), in particular Amendments 45, 47.

This document is available at: www.europarl.europa.eu/supportinganalyses

Contact: poldep-citizens@europarl.europa.eu

Manuscript completed in February 2018

PE 604.942

CATALOGUE: QA-01-18-146-EN-C (paper)
CATALOGUE: QA-01-18-146-EN-N (pdf)
ISBN: 978-92-846-2654-0 (paper)
ISBN: 978-92-846-2655-7 (pdf)
doi:10.2861/047554 (paper)
doi:10.2861/480649 (pdf)