developing platforms themselves enabling active developers to re-use published content, including for publishing purposes.

The evidence across all sectors is that currently UGC is flourishing (as of 2013, 100 hours of video content are uploaded to YouTube every minute). The lack of case law on the issue also suggests that rights holders have so far refrained from preventing its emergence, with notable isolated cases relating to the assertion of moral rights. User groups (in the context of Licences for Europe) have noted however, that, a small portion of UGC may in fact be prevented, and that legal uncertainty as to the possible application of certain exceptions e.g. for quotation, and parody places legal risks on end-users. In Licences for Europe, the positions of stakeholders were too divergent to agree on a common line.

Further details are presented in Annex E.

3.2.1.4. Text and data mining (TDM)

Text and data mining, content mining, data analytics[106] are different terms used to describe increasingly important techniques for the exploration of vast amounts of texts and data (e.g., online journals, web sites, databases etc.). The use of text mining in the field of research has a big potential to foster innovation and bring about economic and societal benefits.[107] Some stakeholders are concerned that the EU might already be losing ground to other regions of the world where TDM is increasingly becoming common practice in scientific research.

Through the use of software or other automated processes, an analysis is made of relevant texts and data in order to obtain new knowledge and insights, patterns and trends. The texts and data used for mining are either freely accessible on the internet or accessible through subscriptions to e.g. journals and periodicals that give access to the databases of publishers.

Usually when applying TDM technologies, a copy is made of the relevant texts and data (e.g. on browser cache memories or in computers' RAM memories or to the hard disk of a computer), prior to the actual analysis. Under copyright law, it is often considered necessary for the making of such copies (even in the case where there is already a lawful access to the relevant text and data), to obtain the authorisation from the right holders [108] in order to mine protected works or other subject matter, unless such authorisation can be implied (e.g. content accessible to general public without restrictions on the internet, open access). Some types of of text and data mining could however fall under the exceptions for non-commercial scientific research in Article 5(3)(a) of the InfoSoc Directive and Articles 6(2)(b) and 9(b) of the Database Directive,[109] which are however optional and have not been implemented in the national laws of all Member States. Some consider that the copies required for text and data mining are covered by the exception for temporary copies in Article 5(1) of the InfoSoc Directive.

It has also been suggested that (certain techniques for) text and data mining do not at all involve copying and therefore are not covered by copyright. None of this is clear, in particular

---

106 For the purpose of the present document, the term "text and data mining" will be used.

107 Big data technologies such as text and data mining have, considered together, the potential to create 250 bn EUR of annual value to Europe's economy (2011 Study of the McKinsey Global institute: Big data –The next frontier for innovation, competition, and productivity)

108 It is common practice in Europe for researchers to contractually transfer their copyrights to publishers

109 Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

since text and data mining does not consist only of a single technique, but can be undertaken in several different ways.[110]

Questions arise as to whether, and to what extent, existing subscriptions (notably to scientific publications) or licence agreements allow for text and data mining. Researchers consider that if a researcher or research institution, or another user, have lawfully acquired access to digital content, including databases, the autorisation to read this content should include the autorisation to mine it.

It has also been argued that it is difficult, onerous and time-consuming to negotiate such agreements with the right holders,[111] and that text and data mining is therefore often undertaken without an explicit permission to do so lawfully. Concerns have also been raised as to the importance of ensuring the establishment of safe and efficient infrastructure for text and data mining, including a secure access to databases used for mining and to control their usage. As an outcome of Licences for Europe, representatives of STM publishers have put forward practical initiatives to facilitate licencing of subscription based content.

Further details are presented in Annex I.

3.2.1.5. Persons with a disability

Digital technology greatly facilitates accessible publishing and today in some Member States 80-90 % of the top titles (books) are simultaneously published in an accessible format for persons with print disabilities.[112] However, it is estimated that at present only 7%[113] to 20%[114] of all titles are available in such formats.[115] In some Member States there are agreements between rights holders and organisations serving the visually impaired for the production, distribution and making available of accessible formats (mainly books), *inter alia* for purposes of education.[116] Such agreements however are not in place in all Member States and only provide access to a fraction of all the works and other subject matter available to persons without disabilities.

The exception for persons with a disability as provided for by Article 5(3)(b) of the InfoSoc Directive[117] is generic and provides little guidance for its implementation. While a number of Member States use the full scope of the exception,[118] others impose limitations as regards the

---

110 However, the CJEU's recent judgment in Innoweb would seem to imply that a licence is required so far as the Database Directive is concerned in the context of comparison websites (see Case C-202/12 Innoweb vs Wegener).

111 The JISC 2012 report "*Value and Benefits of Text Mining to UK Further and Higher Education*" highlights the significant time cost for an individual researcher wishing to mine numerous publications which relates to identifying the right holders and seeking permissions to mine.

112 In the UK 84% of the top 1000 titles in 2012 (source: RNIB), in France close to 90% (source: Exception "handicap" au droit d'auteur et développement de l'offre de publications accessibles à l'ère numerique. Catherine Meyer-Lereculeur, Mai 2013).

113 http://www.rnib.org.uk/professionals/publishing/Pages/publishing_industry.aspx

114 Source : Exception "handicap" au droit d'auteur et développement de l'offre de publications accessibles à l'ère numerique. Catherine Meyer-Lereculeur, Mai 2013

115 These figures represent availability in some but not all accessible formats. Accessible formats include Braille, large print, e-books and audiobooks with special navigation, audio description and closed captioning for films, etc. It is important to distinguish between accessibility from the outset (when a book is created or a film edited in a format that makes it already acessible) from the "assistive solutions" which usually involved the retrofitting of some accessibility features in existing materials. The first one being significantly cheaper that the second one.

116 E.g. the LIA project: http://www.progettolia.it/en

117 The exception may be implemented for any use, for the benefit of people with a disability, that is directly related to the disability and of a non-commercial nature, to the extent required by the specific disability.

118 e.g. Spain, Hungary, Belgium, Poland.

8.13. ANNEX I - TEXT AND DATA MINING

Text and data mining consists of various tools, techniques or technologies for the automated processing of large volumes of texts and data that is often unstructured or not uniformly structured[316]. Mining is undertaken for purposes of e.g., identification and selection of relevant information, retrieval, extraction, interpretation, analysis etc. of such information, and the identification of relationships within/between/across documents and dataset. This allows the miner to obtain new knowledge and insights, patterns and trends. These techniques are increasingly been used across a wide range of sectors and are particularly, although not exclusively, relevant in the field of scientific research.

The large scale use of text and data mining is a relatively new development. Different techniques and software are used for mining. With the evolution of technology, these techniques and software are most probably going to evolve as well. From a legal point of view, the novelty and evolving character of text and data mining techniques raise a number of uncertainties across different fields of law (data protection, fundamental rights, contract law, copyright and database rights, technical standards etc.). As far as copyright (including for databases) is concerned, there is still considerable uncertainty as to the extent to which different text and data mining techniques imply copyright relevant activities or not, and, as the case may be, they are covered by one or more of the exceptions and limitations set out in the EU copyright legal framework.

Besides the legal aspects, practical and technical issues also arise as regards how to ease access to the proprietary infrastructures hosting the content to be mined while safeguarding their stability and security.

Different scenarios may arise. A wide proportion of content (copyright protected or not) currently used as a source for mining is freely accessible on the internet[317] (e.g., blogs, web sites, free sections of online newspapers or magazines, databases, open access scientific journals, etc.). We understand that mining of this content is commonly taking place without any contractual relation between its owner/rightholder and the miners. At the same time, some platform operators have been blocking access to automated analysing of the data on their platforms, including to data provided by third parties (e.g. social networks), for reasons other than copyright.

A different issue arises where content is not freely available online but hosted in proprietary databases/infrastructures (businesses or public authorities databases, subscription based published content such as magazine, newspapers and scientific journals other than open access[318] etc.). If the content owner decides to grant access, it does so by defining conditions and purposes in a contract. Today, scientific articles and research data are considered to be the main source of mining for scientific research purposes. Research institutions or universities typically have access to scientific publications through subscription licences concluded with the publishers. However, such licences usually only authorise the

316 For a description of what text and data mining is, please see chapter 3.2.1.

317 It has been argued in legal literature that content made available on the internet, has been made available with the right holders 'implied consent. This interpretation has been upheld by the German Federal Court of Justice, in the case Abbildung von Kunstwerken als Thumbnails in Suchmaschine [Display of Works of Art as Thumbnails in Search Engine], GRUR, 628 (2010), See "Google and the thumbnail dilemma – "Fair use" in German copyright": http://moritzlaw.osu.edu/students/groups/is/files/2013/08/8-Potzlberger.pdf

318 The growth of open access publications is challenging the traditional subscription model by making scientific publications freely available on-line

reading/consultation of these publications but either do not regulate/authorise or explicitly exclude text and data mining.

When it comes to copyright protected content, the possible need to obtain a specific authorisation to carry out mining (on top of the authorisation to access the content for reading/consultation purposes) depends on a) whether such mining involves a copyright relevant act (in particular an act of reproduction or extraction of data from a database) and b) whether this act may be covered or not by an exception or limitation in the territory where it is carried out.

It is our understanding that current text and data mining techniques usually involve the making of a copy of the relevant texts and data or of parts of them (e.g. on browser cache memories or in computers' RAM memories or to the hard disk of a computer).[319] Copying of copyright protected content constitutes an act of reproduction protected under the rightholders' exclusive rights granted by Article 3 of Directive 2001/29/EC and Article 5 of Directive 96/9/EC. The copying of such texts/data/databases for the purpose of mining may also constitute an act of extraction which is protected by the exclusive *sui generis* right of the maker of a database under Article 8 of Directive 96/9/EC. [320]

Certain acts of reproduction or extractions carried out in the context of text and data could however fall under the exceptions for non-commercial scientific research in Article 5.3 a) of Directive 2001/29/EC and Article 6.2 b) and 9 b) of Directive 96/9/EC. Those articles leave a broad margin of manoeuvre for Member States to adopt, under some conditions, national exceptions allowing the reproduction and extraction of content for the purpose of non-commercial scientific research. If an exception applies, miners do not need to obtain rightholders 'authorisation to engage in those acts. However, the research exceptions are optional and not all Member States have implemented them into national law.

Examples of Member States that have not implemented the exception in Article 5.3 a) of Directive 2001/29/EC are **Denmark**, **Finland** and **Italy**. Other Member States have implemented that exception in a more restrictive way, than provided for in the Directive. Article L. 122-5 of the **French** Copyright Act, limits the use of works for "illustration of research" to "reproduction and presentation of extracts of works".[321]

The **German** copyright act limits the research exception to certain copyright relevant acts, such as the making available of limited parts of a work to e.g., specifically limited circle of persons for their personal scientific research. As regards reproduction, the German act provides that it shall be "permissible to make single copies of a work or to have these made [...] for one's own scientific use if and to the extent that such reproduction is necessary for the purpose and it does not serve a commercial purpose".[322]

The French and German laws do however not contain any obligations to indicate the source.

Article 34 of the **Spanish** Copyright Act also contains an exception for research which is undertaken for non-commercial purposes. It is mandatory to indicate the source of the work.[323]

---

319 An analysis is thereafter made of relevant texts and data through the use of programmed algorithms, software or other automated processes, in order to obtain new knowledge and insights, patterns and trends. The result from the analytical part of the mining would generally be combined, related or integrated with other existing or new information and knowledge

320 See the recent judgment of the CJEU in Case C-202/12 (Innoweb vs Wegener)

321 http://www.culture.gouv.fr/culture/infos-pratiques/droits/exceptions.htm.

322 http://www.gesetze-im-internet.de/englisch_urhg/englisch_urhg.html.

323 https://www.boe.es/buscar/pdf/1996/BOE-A-1996-8930-consolidado.pdf

Moreover, to date no Member States has adopted specific copyright legislation covering text and data mining on the basis of the research exceptions. We are also not aware of any judicial decisions in the Member States touching upon text and data mining, to what extent such activities may be copyright-relevant and whether they could be captured under the research (or other) exceptions laid down by the EU *acquis*.

In June 2013, the **UK** put forward a draft proposal to include a specific exception for text and data mining in its national copyright legislation[324]. The proposal refers to the existing exception in Article 5.3 a) of Directive 2001/29/EC for non-commercial scientific research. In addition to the UK, other Member States (for example France and Ireland) are also discussing the possibility to introduce an exception for text and data mining in their national legislation.

It has also been argued that the mandatory exception to the reproduction right laid down in Article 5.1 b) of Directive 2001/29/EC could apply to at least certain mining techniques. This exception covers temporary acts of reproduction that enable lawful use of a work or other subject-matter, provided that the copies made are transient or incidental. However, it is unclear whether text and data mining would generally fulfil the conditions set out in Article 5.1, since mining techniques usually seem to imply the making of copies which are not temporary and transient.

Market situation

Text and data mining was initially used mostly in the areas of life sciences and drug discovery[325] but is today becoming a common tool also in social sciences, humanities, social media, security, business and marketing and even the legal field. Text and data mining techniques are used on a daily basis not only by researchers but also in business, in particular in the fields of pharmaceuticals, chemistry, abstracting and indexing services, libraries, suppliers of mining tools and services, publishers etc.[326]

Text and data mining was thus initially mostly used in life sciences, with the potential to transform the way scientists use the literature. Some studies indicate that text and data mining can save reading time, information handling time and costs.[327]

Vast amounts of new information and data are produced and put online every day through economic, academic and societal activities.[328] The volumes of such "big data" 1 are predicted to increase at a rate of around 40% per year, and have significant potential economic and societal value.[329]

---

324 http://www.ipo.gov.uk/techreview-data-analysis.pdf

325 Text and data mining constitutes an important tool for the discovery of patterns and relationships in biological and medical research, which is beneficial to the health care sector and to consumers. In this context, the use of text and data mining techniques has already enabled new medical discoveries, e.g., by the linking of existing drugs to new medical applications and by improving human curation. See the response from the British Library to the Independent Review of Intellectual Property and Growth, p. 31, Case D: http://pressandpolicy.bl.uk/imagelibrary/downloadMedia.ashx?MediaDetailsID=886 and http://www.biomedcentral.com/1471-2105/10/326

326 "Journal Article Mining: A research study into Practices, Policies, Plans and Promises", by Eefke Smit and Maurits van der Graaf, 2011, p. 6.

327 "The Value and Benefits of Text Mining", JISC, 2012.

328 "The Value and Benefits of Text Mining", JISC, 2012, p. 3.

329 It was reported in 2011 that if US health care could use big data creatively and effectively to drive efficiency and quality, the potential value from that data could be more than $300 billion in value every year. In Europe, it is argued that government expenditure alone could be reduced by EUR 100 billion a year in operational efficiency improvements alone by using big data. "Big data: The next frontier for

The global research community generates over 1.8-1.9 million new scholarly articles per year.[330] The number of articles published each year and the number of journals have both grown steadily for over two centuries, by about 3% and 3.5% per year respectively. The reason is the equally persistent growth in the number of researchers, which has also grown at about 3% per year and now stands at between 6 and 9 million, depending on definition.[331]

In the field of scientific research, text and data mining facilitates the research process and makes it more efficient, in particular by dramatically speeding up text and data analysis. This increases research efficiency and, as a consequence, the potential to achieve new discoveries. Text and data mining is also an important tool for ensuring, through peer review, the quality and accuracy of research.[332]

Legal uncertainty as regards copyright and text and data mining have come to the fore in particular as regards mining of subscription based content such as scientific journals published under the "traditional" model under which researchers transfer their copyright to STM publishers. Here, the practical question arises as to whether mining should be subject to a specific contractual agreement between publishers and research institutions in addition to the authorisation to access granted through a subscription licence. Currently, it appears that most subscription licences do not include a specific authorisation to text and data mine. Some may explicitly forbid it[333].

In this context, **researchers and research institutions** (such as university libraries) consider that if they have lawfully acquired access to digital content, including databases, the autorisation to read this content should include the autorisation to mine it. In addition, they report high transaction costs mostly due to the necessity for institutions having subscribed to scientific journals to contact a large number of publishers to negotiate and obtain the authorisation to mine their collections[334]. The need to negotiate with each and every publisher[335] also makes the process time-consuming. Researchers have reported cases where they have had to keep ongoing research on hold for weeks or months while waiting for the

---

innovation, competition, and productivity", McKinsey global Institute, 2011, p.2: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

330 "The STM report - An overview of scientific and scholarly journal publishing", 2012, p 5. http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf

331 Around 20% of these are however repeat authors. See, "The STM report - An overview of scientific and scholarly journal publishing", 2012, p 5. http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf

332 Researchers have explained that the peer-review of mining based research involves a repetition of the same mining process as the one undertaken for the research that is being reviewed. In this context, the reviewer needs access to the material on the basis of which the mining was undertaken.

333 See Article "Open Content Mining" by Peter Murray Rust, Diane Cabell and Jennifer C Molloy and slide nr 9 of the following presentation held by a researcher in the Working group on Text and Data Mining in Licences for Europe: http://www.slideshare.net/rossmounce/content-mining

334 The main costs are related to the negotiation of a large amount of licence agreements and also to the setting up of text mining: "The Value and Benefits of Text Mining", JISC 2012, p. 3.

335 An example concerning the PubMed database that contains biomedical literature: in that database there are 587 publishers with more than 1000 papers published each since 2000, see "The STM report - An overview of scientific and scholarly journal publishing", 2012, p. 54. Another example provided by a researcher at the University of Bath is that the 500 most relevant journals for his research are published by 120 different publishers and that the 3 biggest of those publishers combined can provide him with less than 50% of the material to which he needed access: http://www.slideshare.net/rossmounce/content-mining

signing of a licence agreement[336]. Moreover, it has been held that access is often provided only to abstracts of articles[337] and not to the full texts, thus limiting the effectiveness of mining.

Researchers and libraries argue that they are in a position of weakness in negotiations with publishers and that it is difficult to convince the latter to include text and data mining in existing licence agreements[338]. Moreover, in some cases, the benefits of text and data mining can be significantly reduced if not all the relevant literature is captured, i.e., if one of all relevant publishers whose consent is sought for the project refuse access to his content. Finally, research institutions have pointed out that text and data mining should not be limited to non-commercial research[339]

Rightholders, in particular representatives of **STM (scientific, technical & medical) publishers** have held that licensing of text and data mining for scientific purposes is taking place, although they rarely receive requests for an authorisation to use their content for the purpose of text and data mining.[340] The reasons for this could be the potentially high transaction costs described above[341], but also the legal uncertainty surrounding the matter, which could be stimulating the emergence of a "grey market": mining of scientific journals may be actually taking place in a number of cases without it having been specifically licenced with the original subscription.

Even when mining is licensed and takes place, concerns have been reported as to the security and stability of publisher's technical infrastructures hosting the content, due to the intrusive nature of automated processes and mining software (mining techniques usually involve the copying of large quantities of content stored in proprietary databases). In this respect, contractual agreements may be used as a tool to control technical access to proprietary data,

---

336 See slide nr 10 of the following presentation held by a researcher in the Working group on Text and Data Mining in Licences for Europe: http://www.slideshare.net/rossmounce/content-mining

337 See "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles", Blake C, http://www.ncbi.nlm.nih.gov/pubmed/19900574?dopt=Abstract, where the author concludes that the abstracts of articles do in general contain only 8% of the scientific claims and that it therefore is necessary to have access to the full text articles. See also the presentation by Jean-Fred Fontaine "Text and Data Mining for biomedical Research", http://www.slideshare.net/libereurope/the-researcher-perspective-jeanfred-fontaine-mdc-berlin

338 In a study undertaken by publishers, 60% of the seven interviewed publishers replied that they grant researched-focused mining requests in most or all cases. 32% of the seven interviewed publishers replied that they allow text and data mining for all and any purposes *without authorisation needed*, including the 28% that have an open access policy for that. 35% of the seven interviewed publishers replied that they do generally, *upon a request for authorisation*, allow mining in all or the majority of cases, and another 53% said that they allow it in some cases. Again, 53% held that they will decline mining requests if the results can replace or compete with their own products and services. See "Journal Article Mining: A research study into Practices, Policies, Plans and Promises", by Eefke Smit and Maurits van der Graaf, 2011, p. 5.

339 Wellcome Trust, Submission to the UK IPO consultation on copyright, 2012, p. 8 http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtvm054838.pdf; Open Knowledge Foundation, submission to the UK IPO consultation on copyright, 2012 http://science.okfn.org/2012/03/21/response-to-ipo-consultation-on-text-mining-copyright-exception/

340 "Journal Article Mining: A research study into Practices, Policies, Plans and Promises", by Eefke Smit and Maurits van der Graaf, 2011, pp 5 and 31 where only 21% of the seven interviewed publishers responded that they receive more than 10 requests for mining per year, and these are larger publishers.

341 CRA report "Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU – analysis of specific policy options", p. 41.

even independently from profit considerations. Publishers are also concerned that mining may result in the making, and subsequent dissemination, of derivative and/or substitutive products such as summaries or news-clipping based on their publications and are keen to regulate this contractually.

In order to improve the current market situation, representatives of publishers have developed a series of **initiatives aimed at facilitating licensing agreements** for the purpose of text and data mining. In particular, in November 2013, **as an outcome of the "Licences for Europe" stakeholders' dialogue**, a group of STM publishers presented a declaration of commitment covering both contractual and technical initiatives to streamline licences for non-commercial mining of subscription based scientific publications[342].

As reported in this declaration, the signatories have established (and committed to apply) a sample licence clause, to be included in existing subscription agreements (on request or as part of subscription renewal) at no additional cost for the final user authorising text and data mining for non-commercial research purposes A web based click-through licence allowing individual researchers to request this authorisation has also been developed. Technological solutions which could complement the model clause and practically facilitate access to the scientific publications for mining purposes are also being developed. One of the main projects in this respect is the "Prospect" mining hub developed by CrossRef[343]. "Prospect" will allow researchers to access content subscribed by their institution directly in the publisher's infrastructure and facilitate its mining for example through content formatting.

Other initiatives are being carried out at national level. These include work between publishers and rights clearance agents and collecting societies to implement licensing systems to facilitate easy, "one-to-many" rights clearance, such as PLS Clear in the UK[344].

342 http://www.stm-assoc.org/2013_11_11_Text_and_Data_Mining_Declaration.pdf. See also the Commission document "Licences for Europe: ten pledges to bring more content online" http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf

343 http://www.crossref.org/

344 http://www.plsclear.com/