Legal aspects of text mining

Maarten Truyens, Patrick Van Eecke
University of Antwerp
Stadscampus, S.V.121Venusstraat 23, 2000 Antwerpen, Belgium
E-mail: maarten.truyens@uantwerpen.be, patrick.vaneecke@uantwerpen.be

Abstract

Unlike data mining, text mining has received only limited attention in legal circles. Nevertheless, interesting legal stumbling blocks exist, both with respect to the data collection and data sharing phases, due to the strict rules of copyright and database law. Conflicts are particularly likely when content is extracted from commercial databases, and when texts that have a minimal level of creativity are stored in a permanent way. In all circumstances, even with non-commercial research, license agreements and website terms of use can impose further restrictions. Accordingly, only for some delineated areas (very old texts for which copyright expired, legal statutes, texts in the public domain) strong legal certainty can be obtained without case-by-case assessments. As a result, while prior permission is certainly not required in all cases, many researchers tend to err on the side of caution, and seek permission from publishers, institutions and individual authors before including texts in their corpora, although this process can be difficult and very time-consuming. In the United States, the legal assessment is very different, due to the open-ended nature and flexibility offered by the "fair use" doctrine.

Keywords: copyright, database law, case law

1. Introduction

From a legal point of view, *data mining* has already been extensively discussed in the field of data protection, because it is strongly linked with the topics of profiling and behavioural advertising. Conversely, *text mining* has received much less attention, because it has less (direct) privacy impact. Even so, the European Commission recently acknowledged the importance of text mining, and wants to promote its use for scientific research purposes.[1] While attention for text mining is slowly rising[2], legal discussions remain scarce.

2. Applicable legislation

There currently exists no specific legislation about text mining. Relevant rules are mostly found in intellectual property law, contract law and (not further discussed here) data protection legislation. Intellectual property law can be further subdivided into different intellectual property types, among which only copyright and database rights are truly relevant for text mining. Patents (another intellectual property right) can be largely ignored for text mining — at least in Europe, where software, algorithms and business methods cannot be patented. Conversely, other jurisdictions such as the United States do allow such patents, resulting in thousands of software patents and expensive lawsuits against developers.

3. Copyright

Copyright legislation protects "original" texts, sounds or images (not mere ideas) in an automatic way, without any formalities or registration being required. Copyright protection applies worldwide. It acts in a very strict way, which generally prevents copying, modification or publication without the permission of the author — irrespective of the fact that a work would be freely available online. If a work does not meet the originality threshold, it can be freely copied, reused, and distributed without any permission.

"Originality" has a special meaning in copyright legislation (Rosati, 2012). It does not mean that a work must be truly creative, new or beautiful, or require significant efforts to create. Instead, a work needs to be "*the author's own intellectual creation*", and needs to involve *"free and creative choices",* as decided by the EU Court of Justice (CJEU, the highest court in Europe) through a series of recent decisions that have finally brought some uniformity in interpretation across the EU.[3] What these criteria mean in practice is far from clear, although they can be summarized as requiring at least some level of creativity and a certain "personal touch" of the author. Conversely, in situations where no (or only little) maneuvering room exists for an author, the personal touch will not be deemed present. National courts in some countries that previously used deviating standards (particularly the UK) are still catching up to

[1] Communication from the Commission on content in the Digital Single Market, 18 December 2012 (COM(2012) 789 final), see goo.gl/zr1jfZ.

[2] For a general discussion in the context of archiving and digitization, see BORGHI & KARAPAPA (2013). In addition, the Hargreaves review in the United Kingdom (available at www.ipo.gov.uk/types/hargreaves.htm) about the revision of UK intellectual property legislation also touches upon text mining.

[3] *Infopaq v Danske Dagblades Forening*, C-5/08, 16 July 2009 ("Infopaq I"); *Bezpečnostní softwarová asociace v Ministerstvo Kultury,* C-393/09, 22 December 2010; *Premier League*, joint cases C-403/08 and C-429/08, 4 October 2011; *Eva-Maria Painer v Standard VerlagsGmbH e.a.,* C-145/10, 1 December 2011; *Infopaq v Danske Dagblades Forening*, C-302/10, 17 January 2010 ("Infopaq II"); *Football Dataco v Yahoo!,* C-604/10, 1 March 2012.

this new interpretation of the CJEU.

It is already clear that with respect to textual works, at least a combination of words is required, because individual words are not protected.[4] In the famous *Infopaq* decision, the CJEU argued that, depending on its contents, a text fragment of as little as eleven words could be sufficiently original, and thus be protected. Hence, assessing whether an entire text is protected by copyright law requires a case-by-case analysis, because some paragraphs may be original, while others are not. The originality of a text should therefore be assessed both for the text as a whole, and for individual paragraphs, which may or may not meet the originality threshold.

As a general rule of thumb, texts that largely convey mere facts will be less likely to be protected — *e.g.* medical reports, technical manuals, mainstream news events, consumer product reviews, etc. However, even in these relatively "safe" text categories, it is dangerous to assume that none of the texts meet the originality threshold, because some authors can produce very animated texts, even for very dry scientific subject matters. In fact, it is even quite likely that at least some texts do. At the other side of the spectrum, it can be assumed that most traditional "artistic" works (such as novels and poems) will meet the threshold. Obviously, a very large area of uncertainty exists in between.

Only for some types of texts, one can be very certain that no copyright will apply. Depending on the Member State, this is generally the case for official texts (laws and decrees), texts older than seventy years after the author's death (for which the copyright thus expired), and texts explicitly put in the public domain by their author.

Relative certainty also exists for texts that obviously involve no *creative* efforts whatsoever, such as phone listings and other compilations of facts[5]. The same applies to texts that are subject to rigid technical requirements, which leave little room for creative effort — e.g., annotations in corpora, even though these can be very labour-intensive.

Furthermore, depending on the context in which texts will be used, one of the copyright exceptions may apply, which will also prevent the copyright requirements from becoming applicable. In a text mining context, only two exceptions are truly relevant[6].

(a) A first exception, which is implemented in most (but nevertheless differs across those) Member States, is scientific research. Such research must be strictly non-commercial [7] (likely excluding mixed industry academic research, unless a sufficient separation of sub-projects is obtained), and must also indicate the source (including the author's name) of each work used "*unless this turns out to be impossible"*. It is unclear whether such impossibility indeed exists for text mining research, where thousands if not millions of documents are involved. Scientific researchers may also need to take a conservative approach when including texts in their corpora, because the exception only allows the use of a work *"to the extent justified by the non-commercial purpose"*.

(b) A second exception allows to make temporary copies, if such copies are an *"integral and essential"* part of a technological process, have no independent economic significance. This exception tries to reconcile the analogue roots of copyright legislation (where copies usually require permission) with the nature of digital equipment (where any processing involves copying of data). Without this exception, even merely reading a text online would be unlawful, as copies are spontaneously made in the CPU, RAM, router, proxy, browser, etc.

This second exception does not allow the creation of permanent corpora (unless all original elements can be removed — *e.g.,* by only retaining text vectors), because copies must be *temporary*. According to the CJEU, *temporary* does not mean an absolute time limit of a few seconds or minutes, but instead limits the lifetime of a copy to the duration of the "technical process". This exception can therefore be used to create *ad hoc* corpora, to extract non-original snippets of text (*e.g.,* facts, names, numbers, n-grams, etc.), as well as to automatically summarize a text. Still, the exception is not restricted to functions performed entirely by hardware or software, because the CJEU expanded the scope of the exception to also include processes with human intervention, as long as the copies get deleted without human intervention.

Note that this second exception only deals with (temporary) *copying*. It does not allow *publication* of a corpus, for which permission from all individual authors will still be necessary.

Finally, it should be noted that this exception also requires that the copying would enable the "lawful use" of a work, *i.e.* typically an intended use that is simply not protected by copyright law. For example, in the *Premier League* case, the CJEU allowed temporary copies to enable the mere fact of watching a television programme (unlike copying a programme, merely watching it can be performed without any permission, similar to how reading a book is not restricted by copyright either). Applied to corpus creation, the "lawful use" requirement should be met if no copyrightable text fragments end up in the corpus, for example when initial texts are used in order to extract or annotate non-original snippets of text (e.g., facts, names, keywords, numbers, n-grams, etc.), or when initial texts are summarized (by hand or automatically). In all these examples, the initial copyrightable text is no longer used *as such* — instead the ideas, facts, individual words, etc. embedded in it are used.

[4] Individual words may however be protected under a separate intellectual property type (trade marks).

[5] Such texts may however be protected by database legislation, as discussed below.

[6] A third, but only somewhat relevant, exception is the exception to quote from a copyrighted work for purposes "such as" criticism or review. EU Member States are divided on the question whether "such as" results in an *open* list of purposes (*e.g.,* Dutch and Swedish copyright law also allow quotations outside the strict context of criticism and review), or instead an *exhaustive* list of purposes (as is for example the case in Belgium, where the list is limited to criticism, polemics, reviews, education and scientific activities).

[7] The organizational structure and the means of funding of the institution do not preclude the application of the exception. Instead, the non-commercial nature of the research *activity* in question is decisive. See (Hugenholtz & Senftleben, 2011).

4. Database rights

Copyright legislation does not protect compilations of mere facts, because these would not involve any originality. In order to protect investments in fact-based databases created by companies established in (or having strong economic ties with) Europe, a separate intellectual property right for databases was created.

A database will only qualify for protection if substantial investments were made in *obtaining, verifying or presenting* the contents of the database, *i.e.* to search for material, to check such material, and to keep it updated. Investments for creating the individual database entries are not taken into account[8] [9]. Examples of protected databases not only include traditional databases such as medical databases, but also text corpora – and likely even the profiles on a social network – under the condition that the maker of the database is either a national of (or habitually resident in) a European Member State, or a company with its principal place of business within the EEA.

It is prohibited to extract or re-utilize not only a *substantial* part of a protected database, but also insubstantial parts thereof (but only if such is done in a repeated, systematic and unfair[10] way). "Reutilization" should thereby be construed fairly broadly, according to the CJEU[11]. While this remains up for debate, website owners often argue that the server load caused by mining activities are indeed unfair (even though claims have also been raised for other activities, and were not necessarily accepted). Database owners could also argue that text mining without permission undermines their plans to generate license fees specifically for text mining purposes. In light of the growing importance of text mining, such arguments could be very relevant in assessing the normal exploitation and legitimate interests.

Similar to copyright, none of these prohibitions under database law apply in non-commercial scientific research.

5. Contract law

The above analysis assumes that no contract applies. However, except in a few Member States[12], contracts can override most standard legal rules — allowing parties to fine-tune how material can(not) be used. Many licensing contracts of scientific publishers indeed prohibit text or data mining, or require the use of controlled APIs. Negotiating such rights can be a lengthy and complex process, and may result in significant license fee increases.

Also on the open internet, contract law applies through the "terms of use" that are published on many websites. In fact, such terms of use may not only be imposed by one party (e.g., a company or individual publishing a homepage), but may also be simultaneously imposed by several parties[13].

Legal doctrine generally assumes that if these terms of use are sufficiently visible on a website, they will indeed bind the user, even if they are not explicitly accepted by a user by clicking on some button or checkbox. Online terms of use frequently state that the contents of the website can only be used for personal or non-commercial purposes, and/or cannot be downloaded or otherwise permanently stored. Some terms of use also explicitly prohibit "crawling" and "scraping". Copying text fragments of the website will constitute a breach of contract, for which damages can be claimed — even though few case law exists (see Jennings & Yates, 2009). Social networks such as Pinterest[14], Facebook[15] and Tumblr[16] explicitly prohibit both crawling and scraping, while Twitter prohibits scraping but allows crawling on some parts of its site[17]. Anecdotal evidence seems to suggest that social networks are prepared to take legal action towards scraping activities that breach their terms of use[18].

As a result, contractual (and accompanying technical) limitations imposed by scientific publishers may in fact

[8] The following efforts are for example not taken into account: creating or updating the annotations for the individual texts of a corpus; drafting articles in a database with scientific journals; calculating the current index figure for a collection of shares of listed companies.

[9] An interesting question is whether the investments made for annotating the individual texts of the corpus should be taken into account to assess whether the corpus is protected by the sui generis database right. In our opinion, the answer to this question depends on the type of annotation. When annotations are directly embedded in the individual texts (*e.g.,* when tags are inserted to indicate the lexical function of each word), they concern the creation of the individual items, and should therefore be left out of the analysis. If, instead, annotations or similar metadata are created about each individual text *as a whole*, then the investment can be argued to relate to "presenting" the individual items or enriching the overall database layer, and can thus be taken into account.

[10] Article 7.5 of the Database Directive states: *"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."*

[11] In case C-202/12 ("Gaspedaal"), the CJEU held that the concept of re-utilisation covers *any* unauthorized act of distribution to the public of the contents of a protected database or of a substantial part of a database, also covering meta search engines.

[12] In Belgium and Portugal, the copyright exceptions are mandatory law and cannot be derogated from through contracts. Somewhat similarly, in Denmark exceptions can only be derogated from in negotiated agreements (excluding standard agreements such as those accepted through an "I accept" button on the Web).

[13] For example, on a social network, one has to take into account the terms of use of the social network, but possibly also specific licensing terms published by an individual user for the content he or she publishes on the website.

[14] See the Acceptable Use Policy at about.pinterest.com/use/.

[15] See article 3 of the "Statement of Rights and Responsibilities" www.facebook.com/legal/terms, the *robots.txt* file (facebook.com/robots.txt), as well as the Automated Data Collection Terms. Facebook does allow some of the popular search engines to crawl its site.

[16] See section 3 of Tumblr's Terms of Service at www.tumblr.com/policy/en/terms_of_service

[17] Article 8 of the Twitter *Terms of Service*.

[18] See goo.gl/zitW.

constitute the primary stumbling block for text mining in some scientific disciplines.

6. Assessment

Outside the context of non-commercial research, the creation of corpora can be difficult to reconcile with the strict rules of copyright and database law. Conflicts are particularly likely when content is extracted from commercial databases, and when texts that have a minimal level of creativity are stored in a permanent way. Legal rules oppose much less against temporary copies. In all circumstances, even with non-commercial research, license agreements and website terms of use can impose further restrictions. Accordingly, only for some delineated areas (very old texts, legal statutes, texts in the public domain) strong legal certainty can be obtained without case-by-case assessments.

The compatibility between text mining and the traditional principles of EU copyright law is thus a mixed bag. Perhaps the most glaring contrast between current text mining practices and copyright law is the "indexing" performed by search engines such as Google, which store integral copies of web pages into their vast internal databases. Even though various aspects of Google's services have been challenged on the grounds of copyright law (Google News [19], Google's public caching service [20], Google Books [21], Google Image Search[22]), we are not aware of any copyright cases in Europe that deal with the internal indexing process of Google Web Search. Even in those court cases that were aimed at getting (a fair share of) the profits of Google – such as Google News – the focus of the plaintiffs was clearly on Google's front-end, and not its internal processes[23] [24].

In the United States, the legal assessment is very different. First, no database law comparable to the EU exists. Secondly, a "fair use" defence applies that is much more flexible than the strictly limited exceptions in EU law. Under fair use, when a work is either modified or given a new meaning or purpose, its use may be permitted by courts, even if the work would be copied verbatim — effectively serving as a "safety valve" (Leval, 1990) to prevent copyright from becoming an obstacle to scientific progress. Due to its open-ended nature, fair use leads to more flexibility, and can better fit future socio-economic and technological developments without having to constantly update formal legislation. US courts have therefore already allowed the creation of commercial corpora from creative individual texts [25], as well as the mass-digitisation of printed books[26].

The fundamental societal question is whether authors should be allowed to prohibit – and get remunerated for – the use of a work for entirely different purposes. If a teleological point of view is followed, the answer seems negative, because copyright intends to protect the expression, while in a text mining context texts become "raw materials" that acquire their value through the aggregation with many other texts.

The EU thus risks to run behind due to its rigid legal rules: it may for example surprise scientific researchers and engineers that even innocuous aspects of search engines (such as their internal indexes or their display of thumbnails) are difficult to reconcile with the EU rules of copyright.

All hope is not lost, however. Even though many lower courts have given an interpretation to existing exceptions and limitations that is too narrow to respond to new technological developments or business models, some of the highest national courts have instead found creative workarounds — *e.g.,* by arguing that the reuse of copyrighted works should be allowed if it does not harm the authors (similar to how some principles of real estate legislation)[27], or that websites have granted *implicit* licenses when publishing material out in the open[28].

Also, the European Commission recently acknowledged the importance of text mining, and wants to promote its use for scientific research purposes[29] by launching a working group to study the obstacles.[30] Following a public study on the state of intellectual property legislation (Hargreaves, 2011), the UK government

[19] *Copiepresse v Google Inc*, Court of Appeal Brussels, 5 May 2011, English translation available on goo.gl/mIhtlF.

[20] *Ibid.*

[21] See the French case in *Editions du Seuil et autres v Google Inc et France*, Paris District Court, 18 December 2009.

[22] *Vorschaubilder I,* BGH, 29 April 2010, Az. 1 ZR 69/08; *Vorschaubilder II,* BGH, 19 October 2011, Az. 1 ZR 140/10.

[23] In *Copiepresse v Google*, the first court (13 February 2007, *Auteurs &* Media 2007, 107) even stated explicitly (chapter 8) that it was not the copies in the internal cache that were being contested, but their accessibility for the public.

[24] In our opinion, a possible path to take would be that the internal indexing constitutes a *"coutume contra legem"*, i.e. an unwritten practice that is contrary to written law. According to this theory, practices that conflict with written legal rules, but are applied on such a large scale and for such a long time that the average citizen assumes that the practice is lawful, should effectively be considered lawful by courts, in particular when it concerns non-imperative legal rules that can be deviated from through a formal contract. Considering the aforementioned lack of legal interest of authors in the internal indexing process, and the – in information technology terms – very long period of time that search engines have been copying web pages, we are of the opinion that such copying should has indeed become a binding unwritten rule (although it does conflict with the traditional hierarchy of rules). By extension, text mining processes that are sufficiently comparable with typical search engines should then also benefit from this legal defence.

[25] *A.V. v. iParadigms*, 544 F. Supp. 2d 473, 2008.

[26] *Authors Guild Inc v HathiTrust ,* No 11 Civ 6351 (HB), 2012 US Dist.

[27] Spanish Supreme Court April 3, 2012, Sentencia n 172/2012, http://pdfs.wke.es/8/6/1/5/pd0000078615.pdf (see Peguera, 2012).

[28] German Supreme Court (Bundesgerichtshof), *Vorschaubilder I,* BGH, 29 April 2010, Az. 1 ZR 69/08; *Vorschaubilder II,* BGH, 19 October 2011, Az. 1 ZR 140/10. The German Bundesgerichtshof ruled that thumbnails should be allowed on the basis of an "implied license" that was offered by publishing photos online, because the publishers chose not to implement the robots.txt protocol that would have prevented indexation of the site.

[29] Communication from the Commission on content in the Digital Single Market,18 December 2012 (COM(2012) 789 final), available at goo.gl/zr1jfZ.

[30] See goo.gl/ypwizp

launched a bill[31] to allow a person who already has a right to access a work to make copies for non-commercial text mining purposes, without permission. Similar discussions are slowly taking place in countries such as the Netherlands and Australia.

7. Acknowledgements

This article was created in the context of the IWT-SBO project nr 110067. IWT (www.iwt.be) is the Flemish government agency for Innovation by Science and Technology.

8. References

Borghi, M., Karapapa S. (2013). Copyright and Mass Digitization, Oxford, Oxford University Press, pp. 45–-69.

Hargreaves, I., Digital opportunity: A Review of Intellectual Property and Growth, available at *www.ipo.gov.uk/ipreview-finalreport.pdf*

Hugenholtz, P.B., Senftleben M.R.F. (2011), Fair use in Europe. In search of flexibilities., p. 14.

Jennings, F., Yates, J. (2009). Scrapping over data: are the data scrapers' days numbered?, *Journal of Intellectual Property Law & Practice* 4, 2, p. 125.

Leval, P.N. (1990). Towards a Fair Use Standard, *Harvard Law Review,* p. 1105.

Peguera, M. (2012). Copyright Issues Regarding Google Images and Google Cache, in Lopez-Tarruellan, A. (ed.), *Google and the Law – Empirical Approaches to Legal Aspects of Knowledge-Economy Business Models*, The Hague, T.M.C. Asser Press, p. 193.

Rosati, E. (2013). Originality In EU Copyright, Elgar Publishing.

[31] The current state of the implementation of the Hargreaves report can be consulted at www.ipo.gov.uk/types/hargreaves.htm and, specifically with respect to the new data mining exception for non-commercial research, www.ipo.gov.uk/techreview-data-analysis.pdf.