

Free Flow of Data? The Friction Between the Commission’s European Data Economy Initiative and the Proposed Directive on Copyright in the Digital Single Market

Benjamin Raue

Published online: 23 April 2018

© Max Planck Institute for Innovation and Competition, Munich 2018

1 Text and Data Mining as a Reaction to Big Data

The European Union wants to evolve its copyright law into the digital era. The proposed Directive on Copyright in the Digital Single Market¹ is intended to reform Union copyright law, which still mainly relies on the soon to be 17-year-old Directive 2001/29/EC (InfoSocDir).²

The Commission’s proposal is not the major overhaul that some have hoped for and others have feared. Still, it has provoked heated public debates, mainly in two areas: intermediary liability and the press publisher’s right.

Another subject has largely flown beneath the radar of public attention – the proposed text and data mining exception.³ Admittedly, it seems to be a subject for

¹ Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market of 14 September 2016, 2016/0280(COD).

² 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.

³ Although it is subject of an academic discussion, *cf.* Geiger, Frosio and Bulayenko, “The Exception for Text and Data Mining (TDM) in the Proposed Directive on copyright in the Digital Single Market”, Legal Aspects, 2018; Raue, “Das Urheberrecht der digitalen Wissen(schaft)sgesellschaft”, GRUR 2017, p. 11; Spindler, “Text und Data Mining – urheber- und datenschutzrechtliche Fragen”, GRUR 2016, p. 1112; Cocoru and Boehm, “An analytical review of text and data mining practices and approaches in Europe”, 2016; Boulanger, Carbonnel, De Coninck and Langus, “Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU”, 2014; Hargreaves et al., “Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining”, 2014; Triaille, de Meeüs d’Argenteuil and de Francquen, “Study on the legal framework of text and data mining”, 2014; Truyens and van Eecke, “Legal aspects of text mining”, CLSR 30 (2014), pp. 153–170; McDonald, “The Value and Benefits of Text Mining”, 2012.

B. Raue (✉)

Prof. Dr.

Chair for Civil Law, Law of the Information Society and Intellectual Property Law, University of Trier, Trier, Germany

e-mail: iplaw@uni-trier.de

computer nerds and copyright enthusiasts. But it is not. We live in an information society that stores more and more data every day. The amount of today's stored information is estimated to be as large as 20 Zettabytes (= 10^{21} Byte), going up to 160 Zettabytes by 2025. However, without the help of automated search algorithms this data mountain shrinks to a state of electricity on a hard drive.

The computer-assisted search for new knowledge in large quantities of data is called text and data mining. Scientists use these techniques to identify relevant pieces of information in hundreds of thousands of medical papers for a link between genes and a bowel disease, IT specialists to provide search engines, speech recognition or automated translation services, and data journalists to extract publicly relevant information from large amounts of leaked data.

Big data analysis without the help of algorithms is like the search for a needle in a haystack with bare hands – not impossible, but highly coincidental and not sustainable on a large scale. Advancement in knowledge has always been the motor of cultural, societal and economic development. If we want to keep up with other societies, we cannot rely on only gathering a large amount of data; we have to facilitate the access and the use of automated analysis methods.

Thus, there is no alternative to the aim of the European Commission to build a European data economy as part of its Digital Single Market strategy. The Commission aims to make data accessible and reusable by most stakeholders in an optimal way. To achieve that goal it wants to remove barriers that impede the free flow of data and address legal uncertainties created by new data technologies.⁴

2 Copyright Barriers to Text and Data Mining

This said, the Commission and the Member States seem to be very hesitant to adapt copyright law to that strategy, at least in the field of text and data mining.

2.1 Why is Text and Data Mining a Copyright Issue?

At first glance, there is little connection between copyright and text and data mining. Semantic, non-fictional information cannot be copyrighted. That is one of the fundamental principles of copyright law. Therefore, the very process of extracting information does not fall within the domain of copyright.

However, that information is usually contained in copyrightable frameworks such as texts, photos, videos or databases. Whenever a computer processes those frameworks in order to extract the non-protected information it has to create at least temporary reproductions of the copyrighted material. That is the moment when copyright steps in, as Art. 2 InfoSocDir provides the right holder with an exclusive right, even for those temporary reproductions.

⁴ European Commission, “Building a European Data Economy”, COM(2017) 9, pp. 2, 4.

2.2 Why is an Exception Necessary?

Temporary acts of reproduction are exempt from copyright by Art. 5(1) InfoSocDir. Consequently, simple text and data mining activities can be conducted without the consent of the right holder as long as the copyrighted material does not have to be stored for further processing and is automatically deleted by the search algorithm.

For high-quality data analyses, temporary reproductions are usually not sufficient. In some cases, analogue data needs to be digitised before it can be processed. In most cases, the data corpus needs to be normalised, annotated or altered in another way in order to maintain high-quality search results. All those preparatory works usually depend on longer storage periods than Art. 5 InfoSocDir permits. Therefore, text and data mining activities become subject to the right holder's approval, or they need to be covered by a different copyright exception.

A second barrier to text and data mining activities is comprised of contractual restrictions that are imposed especially by the owners of larger databases. The permission to conduct text and data mining is often subject to a further fee and to further restrictions of how to perform the mining exercise and how to proceed with the gathered information. In those cases, the right to read does not include the right to mine. The necessity to identify and negotiate with many different right holders and to implement different research restrictions increases the transaction costs for research activities. That leads to smaller data bodies of poorer quality.

Today's copyright law confronts digital researchers with legal uncertainties even if they have full and legal access to the data body.

2.3 The Proposed Copyright Exception

The Commission has acknowledged those barriers for text and data mining and identified the legal uncertainty as a threat to the Union's competitive position as a research area.⁵ As a consequence, it has proposed a mandatory text and data mining exception. Member States will provide for a copyright exception for reproductions and extractions made by research organisations in order to carry out text and data mining for the purposes of scientific research, provided they have lawful access to the works.

2.4 Its Justification

The exception is justified for three reasons. First, it transfers a core principle of copyright into the digital era. Non-fictional information remains in the public domain. Second, it serves the strong public interest to encourage the generation of new knowledge which would otherwise not exist due to prohibitive transaction costs. Third, it honours another core principle of copyright – the right holders' interest to participate in the economic value of their intellectual property. The exception requires the researcher to have access to the mined material but does not

⁵ Recitals 9 and 10, Proposal for a Directive on copyright in the Digital Single Market.

grant it. The control over access empowers the right holder to charge in the extended use of his or her works and ancillary rights.

2.5 Its Shortcomings

Although the Commission's proposal is to be welcomed in principle, there is one major shortcoming as the exception is limited to non-for-profit research organisations. Commercial research activities are not covered by the exception although they face the same structural problems as non-commercial scientific researchers. High transaction costs and legal uncertainty will either discourage the automated analysis of large amounts of data of different sources, reduce the quality of the research outcome, or lead to a widespread ignorance of copyright law.

The first and second alternative will drive commercial data research to legal systems that provide a more research-friendly environment. This will impede the competitiveness of the European Union and may lead to the relocation of future-orientated workplaces. The third alternative will damage the integrity of the copyright system.

Additionally, the limitation to non-commercial scientific research will create problems for modern investigative journalism that aims to uncover illegal practices and other information of public interest. Nowadays that often depends on the analysis of large amounts of leaked documents. For example, the Panama Papers, which were of a great public interest, consisted of 2.6 Terabytes of data (1 Terabyte = 1,024 Gigabytes). A thorough analysis and the discovery of hidden connections by hand is virtually impossible. Such research activities belong to the core of journalistic work. Article 11(2) of the Charter of Fundamental Rights of the European Union protects the functioning of the press as a public watchdog. The European legislator is called upon to provide for a clear copyright exception covering investigative journalism. Copyright must not be misappropriated to silence critical journalism.

2.6 Proposed Changes in the Legislative Process

Aside from the inclusion of commercial research, there are two proposed alterations to the Commission's proposal in the ongoing legislative process that need to be addressed.

First, the Commission's proposal ensures that the exception is not circumvented by contractual agreements. Although criticised by right holders this provision of the exception is critical to its functioning. One of the major obstacles for the analysis of large quantities of data is the necessity to identify the right holders of the affected works and to conclude agreements with them. If right holders can object to the mining of their copyrighted content or submit it to limiting contractual conditions, then the exception can provide neither legal certainty nor lower transaction costs.

Second, it is being proposed that data bodies need to be deleted after the end of the research activities. There is a legitimate interest of the right holders behind that proposal as they fear the existence of shadow libraries which will take away control of their intellectual property. However, this proposal should be reconsidered.

Especially in the field of scientific research, great effort and large amounts of public money are spent to normalise and annotate the data corpus. It would be a waste of resources if those enriched corpora were not available for later research. The German legislator has found a suitable compromise between the two interests: scientists have to delete their individual copies after the end of their research but may transfer the corpus to a public repository which may store it for later scientific need.⁶ Scientists who have access to the original sources should then be enabled to access the enriched corpus for review or their own research.

⁶ *Cf.* Sec. 60(d) para 3 German Copyright Act; Raue, CR 2017, pp. 656, 660.