

## Should miners go on strike? Legal aspects of text and data mining.

by Pawel Kamocki

PhD candidate

Université Paris Descartes/Westfälische Wilhelms-Universität Münster

Text and data mining (TDM), an automatic process that aims to derive information from datasets, has recently become one of the most important subjects in the debate about the future of IP law. As early as 2001, TDM was listed by MIT among technologies that would change the world in the 21st century. TDM has already allowed some major scientific advances, and has enormous economic potential; the value of the information that can be extracted from public sector data in Europe, for instance, is estimated at 250 million euro.

TDM raises important questions from the point of view of data protection law, but also from the perspective of IP law. Even though raw facts are traditionally beyond the scope of copyright protection, the collection of data may be protected by both copyright (if it meets the originality criterion) and the *sui generis* database right (if it is sufficiently structured and meets the threshold of substantial investment), in which case both a copyright holder and a database producer can forbid the use of their datasets for mining. This general picture becomes even more complicated if we consider that some types of data (such as images or texts) can attract copyright protection on its own.

Thus the 2011 Hargreaves Report was accurate when it stated, '*the law can block (...) text and data mining*'. An opposite claim voiced by copyleft activists, that '*the right to read is the right to mine*' can hardly be defended because unlike human reading, TDM may involve reproduction of the analyzed content. Even though TDM in itself does not seem to be against the *spirit* of IP law, the process cannot fit in the existing European IP law framework. Therefore, in 2012 the British legislator proposed a new copyright limitation allowing data analysis for non-commercial research, causing turmoil among publishers and initiating a EU-wide debate on a specific TDM exception.

The aim of the proposed paper is to argue that even though licensing is definitely not an appropriate solution to allow TDM for non-commercial research purposes (a study shows that researchers should devote 62% of their work time to obtain all the necessary licenses), a new exception may not be an appropriate solution, either. TDM is not a uniform technology, but rather a generic term used to describe various operations that aim to extracting informations from data. The legal status of these operations should be different depending on the type of datasets that are being mined (quantitative data, language data, graphs) and the queries that are being made. Furthermore, at least in some jurisdictions (see: *Mulholland Drive* case in France; Cour de Cassation, 1re Civ, 19.06.2008) this limitation can be interpreted as a shield against infringement claims rather than a sword to fight against technological protection measures. Instead, the paper advocates a broad, technology-neutral research exception as an appropriate solution to the problem.