TEXT & DATA MINING

The Digital Single Market Directive introduces two new exceptions for data mining. This reflects the importance of Big Data and Artificial Intelligence (AI) to the modern European economy.

Whereas the exception aimed at researchers (Art 3) broadly supports data mining in the sector, the general exception (Art 4) for text and data mining (TDM) falls far short of making the European Union (EU) an attractive place for businesses to undertake data analytics or invest in AI.

LIBER believes that, under Article 4, the amount of data available to EU businesses for TDM purposes will continue to be much lower when compared with the US and Japan. Ramifications of this include making the EU less competitive in data and AI markets, and continuing the relatively poor state of publicly available data mining and AI related products and services in less widely spoken European languages.

Key Points

Article 3[1]

This mandatory exception is for research organisations, libraries and other cultural heritage institutions. It allows staff and those attached to these institutions to undertake TDM for commercial or non-commercial purposes. The only exception to this is where the organisation is profit-driven, or a private player would receive preferential access to the results of the data mining. In such instances Article 4 would apply. The derived data produced in the process of TDM can be kept as long as is desirable.

Provisions in contracts that prevent data mining can be ignored, as the exception cannot be overridden by contracts and licences. If data miners face problems with technical measures which prevent mining, these technical protection measures (TPMs) must be removed. This exception is not subject to remuneration.

Article 4

This mandatory exception is for anyone who wishes to undertake data mining. There are no restrictions on the type of organisation undertaking TDM, or whether it is for commercial or non-commercial purposes. The derived data produced in the process of TDM can be kept *"as long as is necessary for the purposes of text and data mining"*.

Contracts can override this exception but where this relates to websites, rightsholders must use *"machine-readable means"* to prevent data mining of websites. If data miners face problems with technical measures which are preventing mining, these technical protection measures (TPMs) must be removed. However, LIBER believes that in such instances rightsholders are most likely to assert via contractual means that their content cannot be mined.

This exception should not be subject to remuneration.

1.Whereas Article 4 allows the mining of all types of works, Article 3 excludes computer programs. We believe this is an oversight that national governments should rectify.

Recommendations For Discussion With National Legislatures

National governments have considerable leeway in introducing the Directive. Libraries should engage in this process. This provision in particular envisages stakeholder dialogue with rightsholders. We recommend that libraries emphasise the following issues:

Resist Third Parties Holding Derived Data (Art 3) - Recital 15 states *"Member States should be free to decide, at national level and after discussions with relevant stakeholders, on further specific arrangements for retaining the copies, including the ability to appoint trusted bodies for the purpose of storing such copies."* This undermines best practice and Research Funder mandates,[2] which allow researchers to decide how and where to store data.

Libraries spend over $7 billion U.S. a year in Europe on content, and if trusted to buy in-copyright materials, they must be trusted to use them.[3] Universities are also examples of best practice when it comes to complying with security concerns over hosting data. They have much experience of hosting security as well as medical data, and of course are GDPR compliant.

Remote Access Required for TDM (Art 3 & 4) - Where an organisation holds analogue materials and wishes to digitise those materials in order to do data mining, data scientists will often not be able to move their data mining technical infrastructure into a library. It is vitally important that member states implement laws that allow remote TDM.

72-Hour Resolution When TPMs Block Access (Art 3 & 4) - Where organisations have invested in data mining and technical protection measures (TPMs) are blocking access, government should be informed of this and access should be granted within 72 hours. Laws should give government the powers to enforce this, including the possibility of fines / compensation being imposed where access is delayed for more than 72 hours.

An Exception To Allow Sharing of Data Mining Results (Art 3 & 4) - Both exceptions are silent on sharing the results of data mining. Under the 2001 Information Society Directive, Member States can introduce an exception for sharing in-copyright materials for scientific research purposes. Doing so is essential to ensure that researchers can share their results, and that science and other forms of research are not undermined.

Robots.Txt Protocol Must Be Used to Prevent Data Mining of Websites (Art 4) - Article 4 allows rightsholders to prevent mining of their websites using machine readable means. Insist that this is done via robots.txt protocol which is an international standard, that allows at the page and item level computer readable prohibitions on copying of in-copyright works. It is important that this solution however does not undermine Article 3.

Key Takeaway

As Member States are now amending copyright law at the national level, **we strongly urge libraries to push for all education, library and research exceptions to be protected from contract override** as is already the case in Belgium, Ireland, Portugal and the UK. All exceptions that appear in copyright law should also be applied to the sui generis database right, so exceptions are the same for both bodies of law.

2.The following funders for example leave it to researchers where to store data as part of their funded data management plans: UKRI, DFG, European Commission.
3.LIBER has used the term "Trusted to Buy – Trusted to Use" to highlight the inappropriateness of any suggestion that libraries are not trusted intermediaries for the purposes of text and data mining.