Article 3 – Text and Data Mining

The ability to analyse information, and to reuse facts and data freely is fundamental to knowledge sharing and everyday life. It should not make a difference whether this happens through simple reading, or with the help of computer programmes that can read (or 'mine') texts and databases to discover ideas and trends. ***Background information overleaf***

Our Ask to improve the TDM exception

- Remove all references to specific beneficiaries, uses or purposes permitted by a TDM exception. This will make for a simpler tool for librarians and others to apply to legally acquired content, and will promote innovation, jobs and growth. It also respects the fundamental principle that the right to read should be the right to mine.
- Remove reference to measures to protect the 'integrity' of databases – this is too vague a term. Ensure that blocking access is an exception, and that use of technological measures to protect works is transparent to the public.

What does the current Commission's Proposal Say? The Commission obliges Member States to ensure that TDM is possible under an exception to copyright. It also provides that neither the terms of contracts, nor the application of technological protection measures should be able to remove the ability to mine, and that it should not be necessary to ask permission from, or pay, rightholders to do this. The exception is however restricted to 'research organisations' carrying out 'scientific research', with only limited possibilities for public-private partnerships. Rightholders are explicitly allowed to implement measures to combat threats to the security and integrity of their databases.

What's Missing?

- TDM, and broader data analysis, does not just take place in universities. As technology advances, it is not just big businesses, but also start-ups, researchers outside of formal institutions, journalists, governments and citizens themselves who are using it.
- The most widely used 'database' for TDM is the Open Internet, not the works of scientific publishers.

By making rules that apply only to universities and scholarly publishers, the proposals, by implication, subject all other text and data mining activities to copyright. Potential users of TDM will be forced to seek licences (difficult to imagine in the case of the Internet), with different laws applying from one country to the next. This would not only kill off many of the activities and jobs that rely on TDM today, but also those into the future.

Start-ups will suffer a competitive disadvantage, while larger firms will look elsewhere to undertake this activity. Citizen scientists and journalists will need to obtain licences to mine content to which they already have legal access, if they have the money. Rightholders will be able to use the poorly defined notion of a threat to the integrity of their databases to restrict access.

Moreover, even within the definition proposed by the Commission, there is no guarantee that libraries are included under the definition of research organisations, meaning that they also will not be able to let users perform text and data mining on materials they have bought or acquired through a license. They will also face confusion in terms of whether a user's TDM work counts as scientific

research, or whether it falls within the definition of a permitted type of public-private partnership.

Finally, while rightholders should be able to take steps to combat piracy, the current provisions will give too much room to stop legitimate uses of TDM. The reference to 'integrity' in particular is overly broad. Europe is behind the US and Japan in terms of research outputs using TDM. But the proposal as stands will leave Europe's businesses, researchers, journalists and citizens worse off than they are now. Scholarly publishers have an important role in facilitating TDM but cannot be given a monopoly.

Background information on Text and Data Mining

Text and data mining (TDM) is the way value is derived from Big Data. It has the potential to speed up research and innovation because computers can process information far more quickly than humans. Governments, journalists, start-ups, technology companies, scientists, and individuals are using it, drawing on a diverse range of sources to take better decisions, uncover stories, identify patterns and advance scientific knowledge. This type of large scale analysis and research does not use or replicate the "creative expression" that copyright is designed to protect. Like the caching exception in EU law[1], TDM is inadvertently caught by copyright because computers make copies in order to function.

The copyright status of TDM in Europe remains uncertain. A significant reason for this is the fact that it is usually necessary to make a copy of works in a format that allows a computer to 'read' and analyse it (for example with pdfs). These copies should not be shared beyond those who have lawful access, or put online, but are a necessary part of the process. The result of this uncertainty is uneven rules, which prevent researchers from working across Europe, if they are allowed to undertake TDM at all.

Many companies, governments, start-ups etc. have been mining web pages they have legal access to for many years. To date this legal activity has been undertaken in a legal "grey zone", but with the Commission and parliamentary debate focussing on the important niche of universities and scientific publishers we are concerned the full picture regarding the business application of Big Data is being lost.

They can also oblige researchers to use their own proprietary APIs. This makes it difficult to treat works from different publishers in the same way, eliminates researchers' ability to choose freely the best software tools for their own research objectives, and raises questions about researchers' privacy. Initiatives such as CrossRef may tackle the first question, but do not answer the others. There have also been cases of researchers and entire universities losing access when legitimate downloading of articles for TDM has been taken for a security risk. As such, the output of European research publications based on TDM has not risen in recent years, while the US, Japan, and increasingly China have seen major increases, representing an increasing share of European research output.

[End of Document]

[1] Article 5(1). Information Society Directive