Ten Things to Know About Text Mining and the Proposed Copyright Directive COM(2016) 593 final

1. Text mining is a process. It is used to derive high-quality information from text using machine – instead of human – processing. Text mining is a subset of the larger activity known as text and data mining, which includes mining of data and non-textual materials. To perform text mining, a copy of the content that is to be mined must be made by the miner. Text mining can be done many ways and with varying degrees of sophistication. At the lowest level, it can be done by scraping websites and converting them to a low fidelity blob (Binary Large Object) of text. At the highest level, it is supported through publisher investment in smart content, ontologies, metadata and format normalisation.

2. STM publishers support, invest in and enable text mining. This investment and support includes producing content in machine ready formats, tagging and enriching content to make it more useful, and creating access through methods such as user interfaces and APIs. Publishers licence the right to text mine to customers in ways that are legally certain and tailored to customers' needs, and in the case of academic users, without additional charge.

3. Text mining is different to search. Search is about merely finding content; mining is about using it. We want publisher content to be found, and ongoing business viability depends on being compensated for its use, whether by humans or machines.

4. STM publishers have already committed to permit text mining of their content at no additional charge to academic/non-profit customers. Publishers have enabled (at their cost, but no additional cost to the user) academic text mining through individual publisher efforts and industry efforts such as Crossref. As such, whilst we consider that an exception extending to those customers is unnecessary, we can work with it.

International Association of Scientific, Technical and Medical Publishers
Prins Willem Alexanderhof 5, Den Haag, 2595BE, NL
www.stm-assoc.org
Registered in The Hague 41200219

5. The exception as currently drafted could have unintended consequences and extend beyond academic/non-profit customers. It could do so in the following ways:

a. The language describing PPPs (Public Private Partnerships) could open the door to potential abuses of the TDM exception by commercial entities. When a researcher is funded in part by a commercial entity with the goal of creating commercial value, that mining should be performed under licence. Likewise, a licence should be required when a non-profit or academic institution is hired to perform text mining for a commercial entity.

b. Lawful access to properly purchased or licenced content is crucial for this exception. Wording must make clear that mining illegal content or mining of copies made under different copyright exceptions by later re-purposing those copies is not permitted. The exception should not allow mining by non-customers who received content by virtue of a loan or as beneficiaries of an exception to copyright.

c. To the operating systems on the publisher platforms, it can be hard to distinguish between a genuine text miner and someone looking to infringe copyright through massive, illegal reproduction of the works. They both bulk download material through automated means. As such, the currently ambiguous language in the directive regarding technical protection creates confusion and puts publisher content at risk. We believe clarification in the directive would be helpful.

d. Copies of the work/subject matter should be deleted as soon as the extraction is completed.

6. There is a significant commercial market for TDM developing and a variety of market-led solutions. Commercial entities are already mining today under licence in the EU, in Switzerland, in the US and globally. These entities include large corporations, start-ups, SMEs, and anyone else who approaches publishers directly or through aggregators such as Copyright Clearance Center. These commercial arrangements vary based on the needs and sophistication of the users.

International Association of Scientific, Technical and Medical Publishers
Prins Willem Alexanderhof 5, Den Haag, 2595BE, NL
www.stm-assoc.org
Registered in The Hague 41200219

7. Today, copyright exceptions for text mining exist in two EU member states: France and the UK. In both, the exceptions are for scientific research and in both, the exceptions are more strictly non-commercial than the draft directive. Japan has a copyright exception which does not extend to databases (including databases of articles, which covers most scientific journals) and with no contract override. Finally, contrary to anti-copyright folklore, copying for purposes of text mining in the US is typically performed pursuant to licence, and is only likely to be deemed fair use where there is no market substitution or market harm to the copyright owner. Use of scientific content for scientific research *is* the market for our members.

8. Concerns have been raised about a situation where an academic researcher uses text mining under an exception, discovers something commercially viable, and then makes that finding commercially available. However, STM's members would be happy if that occured, and would not consider that use to be "commercial" for the purposes of this exception.

9. It should be possible to enter into specific and binding agreements regarding text mining projects. Vague or general contract override language as currently proposed in the directive will make STM's members wary of entering into negotiations with non-profit and academic research institutions for higher level content and services.

10. Publishers routinely create new ways to use and access content for text mining,and indeed other purposes. Technological protection measures (TPMs) are often crucial to this innovation.

International Association of Scientific, Technical and Medical Publishers
Prins Willem Alexanderhof 5, Den Haag, 2595BE, NL
www.stm-assoc.org
Registered in The Hague 41200219