
A II 2 commercial text and data mining (Articles 2(2), 4, 7)

Artificial Intelligence [AI] is a set of technologies that has risen in importance exponentially over the last few years, particularly with respect to digital transformation. A diverse and numerous set of industries and sectors are already benefitting from its value, and that set continues to grow.

Text and data mining [TDM] is a critical tool for use in AI and data analytics solutions -- TDM is a form of technology-enabled analytics that allows for discovering of correlations and identifying useful knowledge from information that rests undiscovered in data sets, large and small, in ways that can be processed and harnessed for a myriad of valuable purposes.

The relationship between AI and TDM is an intimate one. In building AI, TDM of publicly available data, or data for which legitimate access is already available, also plays a crucial role: Access to knowledge is a prerequisite for creating AI and a knowledge economy. In order to understand and use the knowledge stored in data, algorithms must be trained using AI methods. These are neuronal networks that, like the human brain, need to be trained on huge amounts of data, be it just to learn to understand language in general. AI learns about every relevant document, i.e. about every document that is conducive to the respective application scenario.

Which documents are beneficial is recognized by TDM applications and filtered out accordingly. Learning and "seeing" in AI also requires TDM from a technical point of view.

The more relevant data AI "sees" in training, the better it becomes. Only in this way can AI also be created for industrial applications. In the corresponding markets of the future, only those who can access huge amounts of relevant data can be successful. In particular, the freely available Internet offers an adequate database. The possibility of using the freely available data in the network is essential for the development of AI.

Globally, in several jurisdictions, such as, Japan and the United States, researchers are already permitted to carry out TDM for both commercial and non-commercial purposes. There is no doubt that
digital transformation has been slower than expected in Europe – one reason may be because of the status of current legislation, which is causing organisations to undertake TDM and associated AI activity in other jurisdictions, which are perceived to be more AI-friendly.

Against this backdrop, IBM welcomes the recent commercial TDM exception in the DSM Directive as this is an important public signal to make Europe competitive, in order to attract and retain investment and talent in AI. Further, as Germany works towards transposing the DSM Directive into national law, IBM appreciates the opportunity to provide comments to the Federal Ministry.

The German Copyright Act (UrhG) already contains a regulation for TDM (Section 60d UrhG), which, however, is too narrow in many respects. IBM suggests to revisit these existing TDM regulations based on principles as described below:

1. In general, it is a vital next step that any legal and technical hurdles in transposition are kept as low as possible, to ensure that commercial entities (from large corporations with globally distributed locations and personnel, to start ups and SMEs) are able to fully support and engage in commercial take up of TDM in Europe. Further, it is critical that we as lawmakers and policymakers keep at the forefront of our minds, the researchers and technical teams who are on the ground and who will be the key implementers of the TDM provisions of the DSM Directive – it is essential that they have legal and technical certainty moving forward.

2. As such, IBM suggests that language in the transposition related to terms such as “lawful access” and “as long as necessary” in Article 4(1) and (2) respectively, is kept as broad and open as possible, allow for developments in TDM technology and to ensure that researchers and technical teams are not encumbered unnecessarily and arbitrarily in their TDM activities, a view supported e.g., in Recital 18 of the DSM Directive. Reproductions and extractions may be kept for as long as is necessary for the purposes of the TDM (Art. 4 (2)). Thus, the permissible duration of the reproduction is not only based on or limited to the technical TDM process, but deliberately on the purpose of the TDM itself. This is particularly important for the development of AI, because AI is constantly evolving and interacting with the user.

3. Further, with regards to Article 4(3) and particularly, the reservation of rights for online content, there are several issues to be considered. It is important that technical hurdles in any transposition of Article 4(3) are kept to a minimum, because the scale at which TDM and AI solutions function can already be massive and with ongoing technology improvements, the scale is set to only further increase – thus, any technical hurdle/limitation will quickly have ramifications on speed-to-market and progress of AI solutions. Any transposition must be kept broad and flexible enough to accommodate improvements in the advancement of the technology of defacto or standard practices.

With reference to Recital 18, at present, machine readability (as well as cognition) of online terms and conditions is difficult to near impossible for most systems except for the most sophisticated technology. Further, in some instances, it can be difficult to determine where online terms are located/which of the available sets of online terms are applicable.

As such, IBM suggests that, at present, the most feasible method for checking reservation of rights for online content is by using common metadata. Using metadata would overcome the issue of readability as tools to parse metadata can be implemented fairly trivially and economically. Further, use of metadata would also overcome the issue of an indeterminate location for any reservation of rights information, for example, by structuring the metadata according to a predefined format/syntax (e.g., using tags) that can be parsed at a predefined location (e.g., a robots.txt file or equivalent, which is a protocol/format that is used widely by web crawlers and web robots today).
According to Art. 4 (3) any reservation of use must be declared "expressly and appropriately, for example by machine-readable means in the case of content published online". In the English language version of recital 18, appropriateness is specified as follows "In the case of content that has been made publicly available online, it should only be considered appropriate to reserve those rights by the use of machine readable means, including metadata and terms and conditions of a website or a service." This wording makes it clear that a reservation of use for content published online must be declared machine-readable in order to be effective. Otherwise, the reservation would also lead to inconsistencies for online content, because the automated processes require that software can also recognize any reservation. In the German translation of recital 18, this understanding is unfortunately not so clear: "Wurden Inhalte im Internet öffentlich zugänglich gemacht, so sollte es als angemessen erachtet werden, einen Rechtsvorbehalt mit maschinenlesbaren Mitteln auszusprechen." Accordingly, it is important that the German law clarifies that any reservation can only be effectively declared in a machine-readable form.

4. It should also be made clear in the German Copyright Act that such reservations can only be declared with ex nunc effect. This means that reproductions made at an earlier point in time are not affected by the reservation of use, regardless of whether the TDM process has already been completed or not.

5. Further, IBM suggests that for online content, TDM be permitted if prior reasonable efforts have been made to check rights reservations with respect to examining machine readable and understandable data (such as the robots.txt file).

6. Art. 3 and 4 also expressly mention that “database rights” are within the scope of application of the exception according to this Directive. Otherwise, there would be a risk that the TDM exception for content that is freely available in the Internet would run empty if such content would enjoy database right protection. Web pages, whether or not they contain works protected by copyright, may be subject to database protection in their entirety. This special aspect of freely available Internet content also needs to be addressed in the implementation process by adapting the German database right regulations accordingly.

In summary, it is incredibly positive for Europe that commercial TDM will become feasible under the new law. However, as Member States begin to transpose the DSM Directive into national law, we advocate that any legal and technical hurdles are kept as low as possible, in the interests of fostering continued progression and innovation of AI and its related economic and societal benefits for Europe.

IBM is more than happy to be further involved in these continuing discussions.

Terry Ilardi                     Jean-Marc Leclerc
Senior Counsel, IP Policy       Government and Regulatory Affairs Executive
jmleclerc@be.ibm.com

Sirpa Kuisma                    Anita Shaw
IP Counsel                      UK & European Patent Attorney
anita_shaw@uk.ibm.com