March 15th, 2012

STM STATEMENT ON

Text and Data Mining (TDM)

To what extent should copyright law, or laws affecting licensee rights in subscription agreements, be amended expressly to permit consumers and researchers to engage in text and data mining of content to which they subscribe or to which they otherwise have access? To contribute to the public debate on this topic, STM offers the following statements:

- **At present no uniform understanding exists among interested parties as to what TDM is and what it is not.**

This lack of common understanding undermines both agreement and disagreement over the conditions for its acceptability at law and its permissibility under contract. It also hinders the development of best practices and guidelines. STM believes that an essential pre-condition for discussion of legal proposals regarding TDM is that it be clearly defined and for purposes of this paper, we shall use the following **working definition**:

> *A computational process whereby text or datasets are crawled by software that recognises entities, relationships and actions.*

- **Text and Data Mining (TDM) is not an end in itself but a research tool.**

For this reason, STM publishers call for the public debate to recognize that the uses to which TDM is put can and should affect how it is perceived and treated. **TDM undertaken for commercial purposes should be subject to usual rules of commerce** and publishers should be allowed to continue their contribution to a virtuous circle of scientific and economic progress by providing or facilitating customized TDM solutions for commercial purposes as well as by engaging in TDM themselves to improve products and services that, in turn, more effectively support the advancement of science and the economy.

Different from the commercial purposes noted above**, in cases where insufficient market demand exists and TDM is done for non-commercial and purely scholarly purposes**, publishers, as socially responsible organizations, understand that such activities often have the potential to advance the public good without causing economic harm and are predisposed to do what they can in support.

- **TDM solutions are best found in market-based initiatives, like proactive voluntary licensing, that offer faster and more flexible ways to adapt to changing market needs and preferences.**

STM publishers actively provide products and services aimed at supporting the advancement of science through the dissemination and discovery of essential information and key relationships between that information. **STM and its members are currently engaged in evaluating or supporting a variety of initiatives, standards and policy approaches that concern research data and because of STM publishers' heavy involvement is this area they know that issues concerning TDM and published journal articles require special consideration.** The 2011 report from the Publishing Research Consortium titled *Content Mining of Journal Articles* has identified the variety of TDM methods and purposes in which journal publishers and other stakeholders are currently engaged.[1] The PRC Report noted that not all TDM requests come from the general public – many are from abstracting services, commercial entities, and scientific and medical researchers. For researchers and other individuals with a scholarly or non-commercial purpose, it may be indicative that the PRC Report found over 90% of publishing respondents stated that they grant research-focused mining requests.

TDM for commercial purposes involves different evaluation criteria, but increasingly STM publishers are providing licensing options and arrangements for such ends and a fundamental principle of copyright law is that it only protects the "expression of ideas" and not ideas or facts themselves. Publishers and Collective Management Organizations are currently offering a variety of licensing options and permissions, and such offers and solutions should be encouraged and supported. Changes to current copyright law that would mandate extraordinary "rights", particularly when **proposed as copyright exceptions, might not be technically feasible or result in harmful unintended consequences** such as reducing the incentive for innovation in TDM solutions.

***

15 March 2012

ANNEX – DEFINITION OF TDM AND RELATED PROCEDURES AND ISSUES

Text and data mining is a form of explorative data analysis by way of automated, computational and linguistic processes and procedures. TDM differs from and should be distinguished from the following similar procedures:

(i) Information Retrieval (IR): Return of documents or links to documents (ie entire works in a copyright sense) based on a query or search term chosen by the user – a search and report back system.
(ii) Information Extraction (IE): extraction of data from large text or data collection based on known criteria and categories that pre-exist in the databases or are otherwise known to the user seeking to extract information.

[1] See Eefke Smit, Maurits van der Graaf: Journal Article Mining: The Journal Publisher's Perspective, Learned Publishing, 25: 35-46 (Vol 1, 2012), doi:10.1087/20120106; http://www.publishingresearch.net/documents/PRCSmitJAMreport20June2011VersionofRecord.pdf

(iii) Abstracting and summarizing (AS): human-made summary of key claims and findings (abstracts) and summaries (republications in abbreviated or simplified form by humans for humans)
(iv) Automated translation and summarising (ATS): (software tools that produce an automated translation and seek to produce a summary of a text without human intervention.
(v) Query-Response Systems (QR): extracts from Textbases are presented to the user of a query system that best match the user's question.

STM uses the following **working definition** for purposes of this paper:

A computational process whereby text or datasets are crawled by software that recognises entities, relationships and actions.

Characteristics of TDM:

Unlike some of the procedures and tools distinguished from TDM above, TDM seeks the discovery of unknown associations based on categories that will be revealed as a result of computational and linguistic analytical tools.

In this sense, it is a thinking aid assisting the associative thought process of researchers. The need for human interactions to give direction and significance to TDM make it a process requiring central decision making and project management. Frequently TDM is for this reason conducted in-house or within a clearly defined Text or data warehouse.

How to harness TDM to reach its potential:

TDM is rapidly becoming the norm for STM publishers to use text and data mining to improve, add structure and enrich their content offerings.

TDM involves the combination of resources from various stakeholders and in an optimised environment brings together large series of corpuses of standardised or normalised text and data.

The complexity of stakeholder interests and commitments means that no one solution that is imposed externally can be effective, whether by way of national exceptions or otherwise: **solutions must therefore be worked out by multi-stakeholder dialogues which respect stakeholders' interests, are conducted in a transparent manner and aim at models that are simple and scalable.**

***