Text and Data Mining: Building a healthy and sustainable knowledge ecosystem for Europe

What is Text and Data Mining?

Large amounts of new information and data are created every day. Text and data mining (TDM) can help to leverage that information and data to answer specific research questions by uncovering trends and patterns through a process of searching, extracting and analysis. It's not just a simple search, it's like building your own search engine from scratch, and using it to discern new trends within the results that would otherwise not be possible.

TDM has the potential to drive further research and innovation in Europe, but like all good recipes, requires high quality ingredients and the right tools to be successful:

- **Quality content** - Publishers support TDM by providing access to quality content and work with researchers to overcome technical challenges.
- **The right skills** - Content mining requires specific skills to get the most out of it. Uptake depends on education, awareness, tools and infrastructure.
- **Technical solutions** - Software which crawls websites to select and download content can severely affect site performance, so mining access must be managed to make sure that the experience of other users is undamaged.
- **Good user experiences** - Many publishers have developed APIs which provide direct access to data, providing the best possible experience for all users.
- **Collaboration** - TDM requires researchers, librarians, publishers and the wider community to work together. We all have a role to play in making content mining efficient and effective.
- **Investment** - Successful and seamless TDM requires access to quality content and technical expertise. Investment by publishers helps make this possible and sustainable in the long-term.

Publishers support, invest in and enable text and data mining.

Publishers provide access to high quality content and work with researchers to develop new tools, and infrastructure which support TDM.

TDM and the European Digital Single Market strategy[1]

The European Digital Single Market strategy (DSM) presents a vision where digital opportunities are available for all people and businesses, positioning Europe as a world leader in the digital economy. As a digitally forward-thinking European industry, scientific publishers welcome the continued advances that help researchers accelerate new discoveries, whilst contributing to an innovative, competitive and sustainable Europe. Publishers are at the forefront of digital development and as such, support, invest in and enable text and data mining. Publishers are fully committed to encouraging text and data mining of their content at no additional cost to academics and not-for-profit organisations[2].

In order for TDM to successfully contribute to both advancements in science and the future European digital economy a stable and sustainable knowledge ecosystem is required. Ensuring the continued availability of high quality content is vital to achieving these goals.

Text and data mining must be sustainable to take full advantage of its impact on the digital economy.

[1] A Digital Single Market Strategy for Europe COM (2015) 192 final
https://ec.europa.eu/digital-single-market/en/news/digital-single-market-strategy-europe-com2015-192-final

[2] Text and Data Mining for Non-commercial scientific research
http://www.stm-assoc.org/2017_05_10_Text_and_Data_Mining_Declaration.pdf

TDM and the European Union Copyright Proposal

The European Union Commission presented their proposal for copyright reform[3] in September 2016. The proposal will deliver legal certainty to researchers within the EU by enshrining in law TDM activities which are already permitted in several member states and likewise which are already offered by publishers through existing licensing agreements. The proposed new exception provides a guarantee that researchers who are part of public research organisations (and therefore have lawful access to content) can reproduce works for the purpose of text and data mining at no additional cost.

The current Copyright Proposal provides researchers with legal certainty.

This commitment and certainty is already delivered by many publishers, who have enabled (at their own cost) text and data mining through industry initiatives (e.g. CrossRef and Copyright Clearance Centre Mining Services) and the use of solutions based on API access to mineable content.

Publishers enable text and data mining already through a wide variety of industry initiatives and services.

The EU's approach to a pragmatic license-based TDM recognises that it is best served through ruling out unnecessary restrictions rather than imposing broader exceptions to the copyright system which may have unintended consequences.

Wider copyright exceptions are both unnecessary, and likewise have already been noted by the EU as potentially disruptive to a system which is adaptable to the fast evolving TDM environment. The associated Impact Assessment for the EU Copyright Proposal[4] noted that broader exception would:

> "…significantly interfere with the TDM licensing market in the commercial sector."

Likewise, the assessment highlighted that:

> "Commercial companies carrying out scientific research have generally not raised problems with commercial TDM licenses, nor have generally not requested the Commission to take action in this area."

Further Improving the Copyright Proposal

Within Article 3 of the Copyright Proposal 'Text and Data Mining', 'lawful access' is not sufficiently defined. In its current framing, 'lawful access' could cover unforeseen uses such as copying, rented or deposited content or indeed accessing illegally hosted material. Especially within the PPP extension of the exception, it is imperative that "lawful access" be combined with language which confirms the consent of the rightholders. To that end, language from the existing Software Directive[5] of 'acquired lawful access' should be considered for adoption in the Copyright Proposal to provide complete clarity and avoid any unintended adverse interpretations. Finally, to avoid misuse of the source content for TDM, it should be clarified that private partners within the PPP must have valid subscriptions/agreements in place to access the content they wish to mine.

'Lawful access' should be clearly defined to avoid ambiguity and unexpected adverse consequences.

[3] Proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=COM:2016:593:FIN

[4] Impact Assessment on the modernisation of EU Copyright rules
https://ec.europa.eu/digital-single-market/en/news/impact-assessment-modernisation-eu-copyright-rules

[5] Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs
http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32009L0024