Briefing paper:
Text & Data Mining for research and innovation purposes, and its importance

Why is Copyright a barrier to Text and Data Mining?

Text and Data Mining (TDM) is an increasingly common activity for research and innovation purposes. Current copyright legislation in Europe, however, restricts researchers from doing TDM, even though they (or their organisations) have already paid for access to the scientific journal articles they wish to mine.

TDM uses computers to automatically search, filter and interpret large amounts of digital and online content. It automates a process that researchers have done manually for hundreds of years, for example by taking notes. To keep pace with the exponential increase in information, researchers need to mine and analyse content from any source they legally have access to, and which they think is relevant to their research. This content encompasses text and numerical data but TDM is an emerging field that is now being applied to images, video, audio and metadata.

TDM for research purposes is not about gaining illegal access to content - it is about mining content that researchers already have legal access to (e.g. through subscriptions to journals).

Is there growth in the use of TDM?

TDM is now widely embraced in industry - we live in a data-driven economy, with many major companies using it as a significant part of their business. The uptake of this activity for academic research has been far more gradual but it is growing[1]. Searches of Google Scholar by an European Commission-appointed Expert Group led by Prof Hargreaves, reveal exponential growth in mentions of terms related to TDM[2] in the research literature. However, the research sector has been held back by copyright restrictions and the legal uncertainties around TDM's use as well as the practical difficulty of obtaining access and permission to text- and data-mine all relevant material (Boxes 1-3).

Evidence for the demand of TDM and its benefit for academic research

The scale of published research is vast, and it is fragmented across tens of thousands of journals, reports, patents, and more. There are more than 50 million scholarly articles[3] and indexing services are not keeping up with the growth of output[4], meaning that not all the relevant literature is easily

[1] McDonald, D. *et al.* 2012. Value and benefits of text mining. Jisc, London, UK http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

[2] European Commission: Standardisation in the area of innovation and technological development, notably in the field of text and data mining. Report from the Expert Group (2012) http://ec.europa.eu/research/innovation-union/pdf/TDM-report_from_the_expert_group-042014.pdf

[3] Jinha, A. E. (2010) Article 50 million: and estimate of the number of scholarly articles in existence. *Learned Publishing* **23**:pp 258-263 http://dx.doi.org/10.1087/20100308

[4] Larsen, P. O. O. and von Ins, M. 2010. The rate of growth in scientific publication and the decline in coverage provided by science citation index. *Scientometrics* 84:575-603 http://link.springer.com/article/10.1007%2Fs11192-010-0202-z

discoverable[5]. The rate of production of literature is also increasing; MEDLINE indexed 1.13 million biomedical articles published in 2013, with overall output increasing between 5.5% to 8% each year[6]. No single researcher interested in a subset of the literature can feasibly read it all.

Box 1

TDM in the Digital Humanities

Text-mining literature

Toronto scientists Graeme Hirst, Xuan Le and others digitised copies of novels by Agatha Christie, Iris Murdoch and P.D. James and used text mining of vocabulary changes in their later work to detect if there were any indications of Alzheimer's dementia[a]. **This was only possible because of a specific copyright exception under the provision of fair dealing in section 29 of the Canadian Copyright Act. Such text mining is now possible in the UK but is not currently permissible in the EU.**

Citizen scientists helping to research the researchers

Zooniverse have a range of citizen science projects, including for the humanities, which often rely on public access to records that are out of copyright[b]. For example, their ScienceGossip[c,d] project enables any member of the public to describe and tag an image and the illustrator/engraver's name from journals and periodicals in the Biodiversity Heritage library[e], maintained by the Missouri Botanical Garden in the US. A separate international 'Mining Biodiversity' project[f] is developing the algorithms to help mine this information. **The tagging and mining of the images and text is possible only because the content of the Biodiversity Heritage Library is made freely available to the public to access and mine**

a. http://www.cs.toronto.edu/compling/Topics/Graeme-Research/Alzheimers.html and http://ftp.cs.toronto.edu/pub/gh/Lancashire+Hirst-extabs-2009.pdf
b. https://www.zooniverse.org/#humanities
c. http://blog.biodiversitylibrary.org/2015/03/zooniverse-releases-science-gossip.html
d. http://www.sciencegossip.org/#about
e. http://www.biodiversitylibrary.org
f. http://blog.biodiversitylibrary.org/2014/11/crowdsourcing-and-bhl-current-projects.html

Gaining permission from publishers to mine this literature, even when it is for non-commercial purposes, is not practical. The Wellcome Trust estimated that a malaria researcher they fund would spend 62% of their year, at a cost of €25,850, just to obtain the permission from the different journals and publishers involved[7]. Mining is also essential for research such as Systematic Literature Reviews (SLRs) of clinical and pre-clinical studies that can encompass over 1 million research articles[8]. These specialised types of reviews, though, have direct benefits for public health.

Such mining can be done for both non-commercial and commercial research purposes (e.g. by pharmaceutical companies), and often in public-private partnerships (see Box 3). It is badly constrained, however, by copyright restrictions in Europe even though the parties involved have already paid for access to the literature. The use of TDM techniques even has the potential to enable Systematic Literature Reviews that update themselves as and when new relevant studies are published[9,10]. Such an application, while technically feasible, would currently not be allowed under European copyright conditions.

[5] Mounce R. (2015) Dark Research: information content in many modern research papers is not easily discoverable online. PeerJ PrePrints 3:e773v1
[6] MEDLINE trends data http://dan.corlan.net/cgi-bin/medline-trend?Q=
[7] Wellcome Trust Submission to UK IPO consultation on copyright http://www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/wtvm054838.pdf
[8] Shemilt, I., Simon, A., Hollands, G. J., Marteau, T. M., Ogilvie, D., O'Mara-Eves, A., Kelly, M. P., and Thomas, J. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res. Syn. Meth. 5:31-49
[9] Tsafnat G, Dunn A, Glasziou P, Coiera E. The automation of systematic reviews. BMJ. 2013;346:f139
[10] Elliott JH, et al. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. PLoS Med. 2014;11(2):e1001603

Academic researchers are inherently conservative and risk-averse - if there is any hint that their proposed mining work might be construed as 'commercial' it may well put them-off from attempting to do it.

Box 2

The potential of TDM

The following examples illustrate the potential of TDM in different sectors. All depend on mining that is permitted because the content is available in one or more of the following categories: 1) articles are published with an Open Access licence; 2) voluntary sharing of reports or other 'grey literature' online by companies and other organisations; 3) through individual licence agreements with publishers, 4) because authors have permitted TDM to a pre-publication copy of their article or 5) because researchers have gained permission to access a private set of documents. **In each case TDM cannot be applied to the full relevant corpus of information, which limits the extent of the research and the conclusions that can be drawn.**

Text mining to help achieve energy-related environmental sustainability[f]
Mining was applied to corporate sustainability reports made available in a European centralised repository maintained by the Global Reporting Initiative (GRI), which promotes organisations' use of the Sustainability Reporting Framework (GRI, 2013b).

Agro-Know: Data-mining for research & innovation in the agricultural domain[g]
This company helps organisations and researchers find publications, reports, educational resources, projects, people, events and research datasets, and combines these with relevant agricultural content from other pertinent sources. The work is limited to Open Access information or what organisations openly share with Agro-Know.

Text mining to ease the workload in Public Policy Analysis[h]
Researchers Aude Bicquelet and Albert Weale at the London School of Economics are testing the feasibility of using text-mining to analyse large-scale consultations submitted via Internet, such as online consultations and electronic surveys.

PLUTO: Text-mining the literature to make data from evolutionary histories openly available and re-usable[i]
PLUTO is a project at Bath University in the UK to create software to extract data from figures of evolutionary trees within published papers in biology. The scientists involved needed to seek permission from individual publishers to be able to extract this information legally because there was no copyright exception at the time to allow TDM for research purposes. The UK copyright exception for research purposes now puts these researchers at an advantage compared with their EU colleagues.

f. Reuter, N. et al. 2014. Identifying the Role of Information Systems in Achieving Energy-Related Environmental Sustainability Using Text Mining. Proceedings of the 22nd European Conference on Information Systems
g. Agro-Know http://www.agroknow.gr/agroknow/node/1
h. Bicquelet, A. and Weale, A. 2011. Coping with the cornucopia: Can text mining help handle the data deluge in public policy analysis? Policy & Internet 3:1-21.
i. http://gtr.rcuk.ac.uk/project/77B7D1FF-BF3D-49B2-AB1B-F30692D8232C

Evidence for the benefits of, and demand for, TDM for innovation (i.e. outside the academic sector)

The use of mining techniques outside the academic sector often occurs under different names: 'Business Intelligence', 'Data Science', 'Competitive Intelligence', 'ETL' (Extract, Transform, Load), etc. but these all typically involve mining of some form or another. Many successful businesses, from start-ups to SMEs and established large businesses, some in partnership with academic researchers, are using TDM to unlock additional value from published research to facilitate future research (Boxes 1 & 3), but they are hampered in what they can do (legally) if they operate on a commercial basis. Thus

many companies are currently self-restricting their TDM activities to article titles & abstracts only because as commercial entities they do not have permission, nor any legal right, to use TDM on most full-text articles (Box 2).

In terms of a skilled workforce this uncertainty is also limiting training opportunities for a new generation of knowledge workers. The McKinsey 'Big Data' report[11] predicts a significant shortfall of between 140,000 and 190,000 people in the US-alone with sufficient "deep analytical" skills to meet demand in the workplace in 2018.[12].

Why a change in the copyright laws in the EU is needed

The UK has recently introduced a specific copyright exception enabling researchers to make copies for 'text & data analytics'[13] for non-commercial purposes. UK researchers now have a competitive advantage relative to other EU-based researchers. It is expected that UK mining teams will receive additional requests for research collaboration specifically because of this new copyright exception. But both the UK and the rest of the EU remain at a disadvantage compared with the US where Fair Use generally permits TDM for both commercial and non-commercial purposes.

The EU needs to mirror UK policy at a minimum. But the Commission also has an opportunity to develop a more progressive stance by enabling TDM for commercial as well as non-commercial purposes in research and innovation. This will also allow the UK to expand the UK copyright exception to allow TDM for commercial purposes. Ultimately, wholesale copyright reform will be required to enable digital innovation across the European Research Area, as outlined by Julia Reda, MEP[14].

The European Commission itself uses public private partnerships (PPPs) as a major source of funding for research & innovation; more than €6 billion of investment was allocated specifically for PPPs in the Horizon 2020 programme[15]. Restrictions on commercial TDM act as a disincentive to PPPs that would seek to develop and subsequently exploit TDM technology[16] (for example with the European Bioinformatics Institute, Box 3).

We emphasise that researchers from both public and private organisations wishing to mine content already have legitimate access to the content. There is an argument that copyright should not apply at all, as the "copies" made to support mining are ephemeral (as acknowledged in an important US case on transformative use[17]). Crucially, the outputs of TDM are transformational, so do not compete in the market with the original product. Wholesale reproduction of articles for resale is not the aim of TDM and would remain a violation of the rights of copyright holders.

[11] McKinsey Big Data Report 2011
[12] Davenport, TH., and D. J. Patil. 2012 "Data scientist." Harvard Business Review 90: 70-76.
[13] The Copyright and Rights in Performances (Research, Education, Libraries and Archives) UK Regulations 2014
[14] Reda, J. 2015. EU copyright rules maladapted to the Internet
[15] EU industrial leadership gets boost through eight new research partnerships
[16] Reilly, S. 2014. Libraries at the centre of the debate on copyright and text and data mining: the LIBER experience. Paper presented at: IFLA WLIC 2014
[17] Cartoon Network v. CSC Holdings, 2nd Circuit Court of Appeals - http://scholar.google.com/scholar_case?q=Cartoon+Network+v.+CSC+Holdings&hl=en&as_sdt=2,9&case=13763893657469687275&scilh=0

Box 3

Public-Private Partnerships using TDM for research and innovation

The European Bioinformatics Institute (EMBL-EBI) provides services to let researchers perform complex queries on data but is limited in how these can be integrated with the scholarly literature.

EMBL-EBI maintains some of the world's largest molecular databases as well as Europe PMC, a repository of biomedical scholarly literature where much of the knowledge about these data resides. It is imperative that researchers can process information from the literature automatically, to expose relevant knowledge rapidly and effectively, and link it to big data sets. EMBL-EBI work with industry and about 20% of its users are engaged in industrial R&D. EMBL-EBI was able to text-mine Open Access (OA) papers for genes that occur in the same sentence as "Inflammatory Bowel Disease". It took the researchers less than a day - from start to finish - to extract over 6000 relevant sentences from about 800K full text articles[a]. **The UK non-commercial copyright exception now enables EMBL-EBI to do this kind of automated discovery more extensively across millions of articles, rather than operating on the very limited subset of OA articles. However, the extent to which such text-mining at EMBL-EBI can scale and benefit the economy depends on enabling text-mining to all the relevant literature and ensuring that it is available for both commercial and non-commercial research purposes**.

Pfizer and North Carolina State University (US)[b]

A collaboration was arranged between safety researchers at the pharmaceutical company Pfizer and a research team from North Carolina State University in the US to mine ~88,000 articles derived from 4729 journals published over 66 years (from 1945 to 2011) about the safety of 1200 pharmaceutical drugs. This resulted in the generation of 250,000 curated interactions for chemical-induced events (drug reactions). Pfizer limited its TDM to abstracts of scholarly articles. Additional analysis of the full text of some papers was done by NCSU biocurators via their institutional library subscriptions. Had Pfizer had the ability to mine the full text of all relevant papers the exercise would have had data from thousands more articles with which to inform the academic researchers. **Text and data mining for commercial purposes is not currently permitted within the EU (no copyright exception) or even in the UK (no exception for commercial TDM)**.

Conference on TDM bringing together industry and academic research scientists (Germany)

A recent conference was held in Cologne about the application of TDM for research purposes in both commercial and academic settings: "From Big Data to Smart Knowledge – Text and Data Mining in Science and Economy" task or decision https://textmining.congressbuero.de/home . **This demonstrates the increasing demand for PPPs around pre-competitive TDM for research purposes.**

Open PHACTS: reducing pre-competitive barriers to drug discovery with TDM[c]

Open PHACTS is a software platform built through collaboration between major academic and commercial organisations involved in drug discovery, which integrates and links data across multiple publicly available databases to easily see the relationships between compounds, targets, pathways, diseases and tissues. **Current copyright restrictions on the mining of literature limits the capacity of this platform to scale up its operations.**

Patented technology originating from the Oak Ridge National Laboratory (US), that helps to reduce information overload[d]

The US Oak Ridge National Laboratory's Computational Data Analytics Group's has created a text-mining system to discover, recommend and visually represent meaningful information from raw data across thousands of documents. This US-based academic work has resulted in four issued and four pending patents, several commercial licenses, a spin off company, an R&D 100 Award, and scores of peer reviewed research publications. **The ability of academic institutions to create the same sort of innovative technology is hampered by current copyright legislation in Europe[e].**

a. Jo McEntyre, Head of Literature Services, Europe PMC

b. Davis, Allan Peter, Thomas C. Wiegers, Phoebe M. Roberts, Benjamin L. King, Jean M. Lay, Kelley Lennon-Hopkins, Daniela Sciaky, et al. "A CTD-Pfizer Collaboration: Manual Curation of 88,000 Scientific Articles Text Mined for Drug-Disease and Drug-Phenotype Interactions." Database: The Journal of Biological Databases and Curation 2013 (2013): bat080. doi:10.1093/database/bat080. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842776/

c. http://www.openphacts.org/

d. ORNL's Text Mining Technology

e. http://cda.ornl.gov/piranha.shtml

Copyright exceptions will enable TDM to scale

Research is fragmented across more than 28000 peer-reviewed journals from around 10,000 distinct publishers[18] [19]. It is a long-tail problem that additional licensing agreements can't solve: so far the CrossRef Text and Data Mining pilot covers only 9 different publishers[20] and it only facilitates text mining. Some publishers are actively working to make their content more discoverable, such as Springer[21] who have recently granted text- and data-mining rights to their subscribed content to researchers via their institutions, provided the purpose is non-commercial research. Other publishers provide permission to mine the literature for research purposes through individual licence negotiations. These examples highlight the severe fragmentation that exists.

This fragmentation is both technical and legal. Whilst mechanisms that enhance interoperability for discovery are welcome, any 'click-through' licensing for access is cumbersome, and few researchers will have the time and resources to consider each and every different licensing agreement. This becomes impractical at scale – where the real benefits of TDM arise – especially where the content required comes from multiple sources. Content relevant to research is not restricted to the scholarly literature or to text. Canadian language scholars, for example, have applied text mining to the vocabulary used by Agatha Christie and other novelists to test the hypothesis that these authors developed Alzheimer's dementia in their later works (Box 2). As a consequence of these issues, licensing agreements will be unable to keep pace with demand.

Claims are frequently made that TDM could create a strain on publisher websites, although these claims seem at odds with other claims from publishers that 'there is little demand' for TDM. Evidence from PLOS (provided by C. Neylon) shows that the 'strain' imposed on publisher websites by text mining is negligible compared to other normal fluctuations in traffic for any significant website, such as a spike in traffic from social media platforms like Reddit. Traffic associated with text mining is also negligible compared to risks faced by all web sites such as distributed denial of service attacks (DDOS)[22].

SPARC Europe
www.sparceurope.org info@sparceurope.org
98 Watermanstraat, Apeldoorn, The Netherlands

[18] Morris, S. 2007. Mapping the journal publishing landscape: how much do we know? Learned Publishing 20:299-310
[19] Ware & Mabe 2012. STM Report http://www.stm-assoc.org/industry-statistics/the-stm-report/
[20] CrossRef Text and Data Mining Services
[21] Springer's text- and data-mining policy http://www.springer.com/gp/rights-permissions/springer-s-text-and-data-mining-policy/29056
[22] Neylon, Cameron. "Best Practice in Enabling Content Mining |." *PLOS Opens*, March 9, 2014. http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/.