CENL-FEP Statement on Text and Data Mining

1. Introduction

The emerging importance of Text and Data Mining

Text and Data Mining (TDM) is "the computer-based process of deriving or organising information from text or data. It works by copying large quantities of material, extracting the data, and recombining it to identify patterns, trends and hypotheses or by providing the means to organise the information mined. "[1] It is important that the content can be copied into a secure digital space so that it can be tagged and re-formatted for consistent analysis to take place.

The benefits of TDM include: very efficient retrieval of information from large sets of text; the discovery of new knowledge; the creation of new areas of research; and the testing and correcting of traditional research.[2] Examples in the medical field include the identification of new medical treatments, of previously undiagnosed side-effects to medication, and of connections between previously unlinked ailments.[3] It is important that TDM is not seen as simply a form of search engine for data: it is a more powerful and nuanced form of data analysis than that, involving a series of more sophisticated operations than may be first realised.

TDM is of importance to authors, publishers, libraries, government, researchers, companies and the wider public. TDM has long been facilitated by digitisation of content: by publishers, digitising their own content, and by research libraries, including national libraries, digitising out-of-copyright content in their care. In-copyright e-only materials also have great potential for TDM. National libraries are perhaps uniquely in a position to apply TDM to very large corpora of texts, produced originally in multiple and diverse ways.

Recently, publishers have developed fast-track ways of permissions-based clearance for use of their content for TDM, such as the CrossRef TDM service[4]. In the legal environment, the European

[1] UK Government, Intellectual Property Office, Text Mining and Data Analytics in *Call for Evidence* Responses, 2014; http://www.ipo.gov.uk/ipreview-doc-t.pdf

[2] Diane McDonald and Ursula Kelly, JISC, *Value and benefits of text mining*, 2012. https://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining

[3] Min Song, "Opinion: Text Mining in the Clinic", in *The Scientist,* April 1, 2013, http://www.the-scientist.com/?articles.view/articleNo/34820/title/Opinion--Text-Mining-in-the-Clinic/

[4] http://tdmsupport.crossref.org/

Commission is seeking to address TDM as part of the reform of EU copyright[5] and has published a *Proposal for a Directive…on Copyright in the Digital Single Market* (COM(2016) 593 final).

In the belief that TDM is still relatively under-used, an EC-initiated project FutureTDM has now reported with the aim "to improve uptake of text and data mining in the EU by actively engaging with stakeholders such as researchers, developers, publishers and SME's."[6] In the UK, France and Germany there is already a copyright exception for non-commercial use of TDM for those researchers with legal access to the content.[7]

About CENL-FEP

The joint committee of the Conference of European National Libraries and the Federation of European Publishers (CENL-FEP) meets twice a year to discuss and exchange information of mutual interest, including deposit practices and legislation. It seeks to clarify, define and agree areas in common.

The Conference of European National Libraries has the aim of increasing and reinforcing the role of national libraries in Europe, in particular in respect to their responsibilities for maintaining national cultural heritage and ensuring the accessibility of knowledge in that field. Members of CENL are the national librarians of all Member States of the Council of Europe, 49 members from 46 European countries. The Council of Europe is a separate and distinct organisation from the European Union.

The Federation of European Publishers is an independent, non-commercial umbrella association representing 29 national associations of publishers of books, learning materials, and/or journals in a range of media, in Europe.

Both CENL and FEP are membership organisations, and cannot speak for sectors they do not represent (higher education in the case of CENL, for example; and newspaper publishers in the case of FEP, for example).

[5]European Commission,"Commission takes first steps to broaden access to online content and outlines its vision to modernise EU copyright rules", Press release, 9th Dec 2015; http://europa.eu/rapid/press-release_IP-15-6261_en.htm

[6] http://www.futuretdm.eu/news/about-futuretdm/

[7] UK: The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014 http://www.legislation.gov.uk/uksi/2014/1372/made; Germany: https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/GesetzBeschlussBT_UrhWissG.pdf?__blob=publicationFile&v=1; France: https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte

2. CENL-FEP Statement

This statement by the CENL-FEP working group affirms common ground between the publishing and national library communities represented by the group. It sets out at a high level that common ground, and it identifies areas which are more problematic or where further exploration of the issues is needed.

2.1 Common Ground

CENL-FEP recognises that TDM has potential benefits for all.

CENL-FEP recognises that facilitating TDM involves considerable investment and maintenance from national libraries and publishers.

CENL-FEP recognises that TDM is appropriate only for those with legal access to the content.

CENL-FEP recognises that, because TDM involves the copying of content into a digital space, there is a need for copies made under TDM to be protected against illegal exploitation for revenue purposes and potential violations of privacy. Security of data and networks is paramount to publishers and national libraries alike.

CENL-FEP recognises that TDM does not involve copyright uses other than the act of reproduction.

CENL-FEP recognises that there are corpora, often held in national libraries, out-of-copyright works and sound collections for example, beyond the remit of the CENL-FEP group, but which are still vital to the understanding of TDM's potential. Similarly, publishers do not always have databases of structured content.

TDM depends on the deployment of technical developments for access and standardization of content which currently most publishers cannot afford and which currently are beyond the normal business as usual processes of national libraries.

2.2 Areas for further exploration

There are some areas in relation to TDM which are more problematic or where legal and practical implications appear to CENLFEP to warrant further exploration.

- The nature of "legal access"/"legally acquired access" in respect to TDM applied to digital content and has to be properly defined.

- National libraries might act as trusted third parties to facilitate TDM for researchers using publishers' content.
  CENL-FEP seeks to explore how to understand the costs of TDM, and mitigates to reduce these costs.
  CENL-FEP seeks to explore further what using TDM for 'non-commercial purposes' might entail.

- CENL-FEP seeks to encourage the better understanding of TDM benefits to key stake-holders.

2.4 Conclusion

Publishers and national libraries recognise the enormous potential benefits of Text and Data Mining. As a working group of representatives of key stakeholders in this field CENL-FEP affirms substantial areas of common ground between publishers and national libraries, and seeks to continue the dialogue on questions where there is much more to explore.