Research Consulting

# Text and data mining in higher education and public research


An analysis of case studies from the United Kingdom and France

PREPARED ON BEHALF OF THE ADBU

December 2016

Association des Directeurs & personnels de direction des Bibliothèques Universitaires et de la Documentation

**"Text and data mining in higher education and public research"**

*Report commissioned by:*

Association des Directeurs & personnels de direction des Bibliothèques Universitaires et de la Documentation (ABDU)

www.adbu.fr
Contact: Julien Roche, vp@adbu.fr

*Report authors:*

Rob Johnson, Olga Fernholz, Mattia Fosci
www.research-consulting.com
Contact: rob.johnson@research-consulting.com

Report dated: December 2016

# Executive Summary

## Background

This study uses case studies from researchers in the UK and France to assess the value of a copyright exception for text and data mining, and identify the steps needed to realise its potential. It was commissioned by the ADBU, the French association of directors and senior staff in university and research libraries, and delivered by Research Consulting, a UK consultancy specialising in the management and dissemination of research.

## What is TDM?

*'TDM is any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations.'* (European Commission)

More data has been generated by people and machines over the last two years than in the whole history of mankind. We are seeing unprecedented growth in the volume of structured and unstructured data available on the internet and elsewhere - including more than 2.4 million scientific articles published every year. For researchers, reviewing and building on this increasing body of knowledge has become ever more difficult. Text and data mining (TDM) allows them to make sense of this data deluge, employing advanced software to analyse articles and digital information whose volume would be impossible to scrutinise manually.

## TDM magnifies the reach and efficiency of scientific research

Previous studies have indicated that TDM could:

- Increase coverage of domain knowledge in systems biology by a factor of 4
- Identify relevant public policy studies with only 25% of the usual manual work
- Improve productivity in the curation of biomedical literature by 50%
- Identify the top healthcare papers among 1.3 million publications per annum – a process that would 2-3 years to complete manually
- Accelerate drug discovery, reducing the 10-12 year average timeframe from discovery to market

TDM is already being used to address key research challenges, including:

- Improving understanding of climate change
- Driving progress in Alzheimer's research
- Identifying new health hazards
- Helping to predict epidemics
- Using crime statistics to inform policing

## Enabling TDM

### Achieving legal clarity

In order to mine them, materials must be accessed, copied, stored, and sometimes converted into new formats. As such, enabling TDM involves reviewing various regulatory regimes - most notably copyright, but also database and contract law. Many believe that Europe has been slow to take action in these areas and that, as a result, it has been overtaken by Asia as the world's leading centre for academic TDM research. In the US and several other countries, the 'fair use' doctrine gives researchers an 'affirmative defence' to charges of copyright infringement when undertaking TDM.

Until recently, European researchers have lacked similar enabling legislation - but this is now being addressed through exceptions to copyright. The United Kingdom adopted a copyright exception 2014, and the 'Loi pour une République numérique' introduced a similar exception into French law in September 2016. The European Commission has issued draft proposals for an exception which would apply to all member states.

Our study found that the fear of inadvertently breaching copyright law makes European researchers reluctant to use TDM. Copyright exceptions improve matters, but researchers still face legal uncertainties:

- The burden of proof is on the researcher and their institution to show they meet the exception criteria, but there is a lack of expert advice and guidance to assist them
- Researchers in the UK and France worry that their research might not qualify as non-commercial, meaning they cannot rely on an exception
- The French exception applies only to texts included in or associated with scientific writings – it doesn't enable mining of the web or social media
- Only designated organisations are able to conserve and communicate copies arising from TDM in France

### Delivering access to content

Researchers cite difficulties in securing access to published content, and often have to scale back their research questions as a result. Content can be sourced through bulk downloads or web crawling, but in each case researchers may encounter barriers:

- **Technical protection measures (TPMs).** Publishers may impose limits on the speed or volume of downloads to ensure that mining does not corrupt or reduce the quality of their service.
- **Crawler traps.** Publishers introduce web pages that look like an academic paper to an automated downloading programme and block access to the rest of the publisher's content when mining software is detected.
- **Restricted access to APIs.** Application Programming Interfaces (APIs) can be used by researchers to acquire content in machine-readable form. However, not all publishers make an API available, others require users to sign additional agreements before access is allowed, and varying data formats between publishers make aggregation difficult. Worse still, researchers find that some APIs fail to return all relevant documents in response to a programmatic query.

### Developing infrastructure

A copyright exception may discourage publishers and content providers from developing services and enriching content for TDM purposes, so support must also come from public sources. France's investment in the ISTEX project (www.istex.fr) has created a critical mass of content, infrastructure and expertise that bodes well for the future. Further development is now needed to:

- Improve the technologies used to aggregate, normalise, interrogate and preserve TDM materials.
- Encourage wider use of ISTEX for TDM purposes, and improve the availability of recent content, which is crucial to fast-moving research fields.
- Develop user-friendly cloud-based services, which can be used by researchers with minimal technical expertise.

### Enhancing skills and support

TDM requires a high level of digital literacy. Text-mining experts, computer science departments and libraries can all play a role in supporting researchers in acquiring the necessary skills. Improved levels of collaboration between text-mining experts and domain scientists will be necessary as TDM evolves.

### Providing incentives

TDM faces many of the same cultural challenges as the wider move to open science, including:

- lack of awareness;
- the strength and variety of disciplinary cultures;
- the investment of effort needed to gain the necessary skills and conduct the first experiments;
- a lack of widely-available and easy-to-use infrastructure and tools;
- doubts as to the likely value of the results;
- limited funding for data-sharing and curation.

Greater financial and reputational incentives are needed to support the wider adoption of TDM within the research community.

## Conclusion - the current state of TDM

TDM has enormous potential to speed up public research, and to deliver wider economic and societal benefits. The introduction of a copyright exception helps to place European researchers on a level playing field with those in Asia and North America. Existing TDM practitioners in the UK are already seeing the benefits, but overall uptake remains low. More must be done for TDM to become widely used by researchers in both the UK and France.

Our study makes clear that changes to copyright law must be accompanied by improvements in access, infrastructure, skills and incentives for TDM. Strong leadership and additional investment are now needed to create an environment in which TDM can truly flourish. Academic libraries should be at the forefront of this change. Concrete actions which can be taken by them and other stakeholders to help realise the potential of TDM are outlined on the following page.

## Securing the future of TDM

### The role of libraries

Libraries and library consortia have a key role to play in supporting and enabling TDM. Our study identifies five key areas where action can be taken by the library community:

1. Implement mechanisms to monitor researchers' experiences
2. Develop case studies and guidance
3. Secure the support of the national library
4. Incorporate TDM clauses into model licence agreements
5. Invest in dedicated TDM support services for researchers

### Actions for other stakeholders

The creation of an enabling environment for TDM will also require intervention by a wide range of other stakeholders, as outlined below.

#### Legislators

1. Communicate legal provisions for TDM with certainty and clarity
2. Clarify the exception's scope where public researchers collaborate with commercial partners
3. Monitor the interaction of the copyright exception with digital rights management (DRM), licensing and other relevant legal regimes

#### Research leaders and institutional managers

1. Communicate the benefits of TDM as a legitimate research technique
2. Invest in the development of library services to support TDM
3. Explore opportunities for knowledge exchange in TDM with commercial partners

#### Research funders and policy makers

1. Invest in the infrastructure needed to support TDM
2. Identify or create a national forum to address challenges over access
3. Consider the needs of TDM researchers as part of the move to open science

#### Publishers and infrastructure providers

1. Pursue the development of cloud services for TDM
2. Develop arrangements to streamline access to all forms of copyrighted material for mining purposes
3. Adopt open standards and harmonised data formats

# Contents

# 1 Introduction

*Text and data mining has transformative potential for public research. Whether it delivers on its promise depends on an enabling copyright regime. This study, commissioned by the ADBU, uses case studies from researchers in the UK and France to assess the value of a copyright exception for TDM, and identify the steps needed to realise its potential.*

## 1.1 Background

This study was commissioned by the ADBU, the French association of directors and senior staff in university and research libraries[1].

Founded in 1974, the ADBU's mission is to promote and develop record-keeping and libraries in higher education and public research. The ADBU has been a longstanding advocate of an exception to copyright for text and data mining purposes.

The work was delivered by Research Consulting Limited, a UK consultancy specialising in the management, dissemination and commercialisation of research.

## 1.2 Terms of reference

The purpose of the study was to obtain the views of active researchers and practitioners of text and data mining on the legal, organisational and practical implications of its use in public research.

The study aims to provide:

- A comparative analysis of the regulatory regimes for TDM in France and the UK, placing these in a European and global context
- Emerging indications of TDM's potential to increase research efficiency
- Case studies illustrating the barriers and enablers to the use of TDM within public research in the UK and France
- Recommendations on realising the potential of TDM in public research, with particular reference to the role which can be played by academic libraries.

[1] Association des directeurs et personnels de direction des bibliothèques universitaires et de la documentation

## 1.3 Report structure

The report is organised in two parts. **Part A** summarises the context in which text and data mining (TDM) is undertaken across relevant jurisdictions. First, it defines TDM, explains why and how it has emerged as a useful research technique, and briefly considers the factors influencing its uptake (section 2). Then it presents a summary analysis of the regulations affecting copyright and TDM use in the UK, France and the European Union (EU) (section 3). It looks at both the legal regime in operation in the past and the legal changes recently enacted or currently under discussion, giving some consideration to the practical implications that these would have on researchers' ability to exploit TDM.

**Part B** uses a range of case studies of researchers and TDM practitioners to illustrate current practice in the mining of scholarly publications in France and the United Kingdom, and assesses the relevance of a legal copyright exception for this purpose. Both the UK and France have now enacted an exception to copyright for TDM in non-commercial research, and researchers in both countries face many common challenges. Our work indicates that action is needed in five key areas for TDM to become widespread:

- Achieving legal clarity (section 4)
- Delivering access to content (section 5)
- Developing technical infrastructure (section 6)
- Enhancing skills and support (section 7)
- Providing funding and incentives (section 8)

Using illustrative case studies, each section explores the challenges faced by researchers in the considered area and makes suggestions on how these might be overcome. The section concludes by summarising the current situation, and making recommendations to academic libraries and other stakeholders for improving uptake of TDM.

The report contains three appendices: a list of interviewees and contributors to the study (Appendix A), a section detailing the EC, UK and French exceptions to copyright for TDM (Appendix B) and a list of abbreviations and glossary of terms used in the report (Appendix C).

## 1.4 Methodology

The following approach was followed in preparation of this report:

### *Step 1. Desk-based review of relevant literature*

We reviewed the literature on TDM use in the UK, France and globally, taking into consideration both peer-reviewed texts and 'grey literature'. The exercise also clarified the regulatory framework governing TDM and copyright in the selected countries, as well as at a European level.

### *Step 2. Case study identification*

We then sought to explore key issues, common challenges and approaches to TDM through a qualitative analysis of individuals and organisations involved in using or supporting TDM for public research in the UK and France. The analysis is presented as a set of case studies. Contributors were

selected using 'snowball sampling'[2], whereby a small initial group of interviewees provide referrals for others, and so on. The initial set of interviewees was identified from our network of contacts in UK, recommendations supplied by the ADBU in France, the results of the literature review, and web-based searches.

In total, 70 potential contributors were identified over the course of our work. 55 of these individuals were approached formally to request their input, and 25 have been interviewed. Several others provided referrals and written contributions, while the remainder declined to participate, mostly due to a lack of direct involvement in the text-mining of copyrighted materials.

### *Step 3. Case study development*

Interviews with sampled individuals were conducted remotely and lasted about one hour each. Wherever possible, conversations were recorded to ensure accuracy. Full transcripts were then prepared and a summary analysis was used as a basis for the case study preparation. Interview data were cross-checked against desk-based research and complemented by resources provided by the interviewees themselves.

### *Step 4. Validation of findings*

A review and validation process was used to ensure accuracy and balance. Each case study was shared, in draft form, with study participants and feedback solicited. Similarly, feedback on a draft version of the findings summarised in sections 3, 4 and 5 of this report was also sought from study participants. Emphasis was placed on ensuring that the findings were reported in a fair, balanced and accurate way and that no undue conclusions were inferred from the data.

## 1.5 Limitations of the study

This study is subject to a number of limitations:

- 'Snowball sampling' (also known as chain sampling) does not guarantee representativeness of the sample because the researchers have incomplete knowledge of the distribution of the overall population vis-à-vis that of the sample. It is therefore difficult to make inferences about populations based on the obtained sample.
- The challenge of generality applies to case study analysis as a whole, in that generalising from a relatively small sample of organisations risks overlooking or, conversely, overplaying issues affecting the wider population.
- Moreover, case studies are based on the observations and perceptions of one or few individuals: this may affect data quality by presenting implicit risks of bias.
- Ethical and privacy issues related to TDM fall outside the scope of this study, but have been widely explored in previous literature on the topic[3].

[2] Snowball sampling (n.d)
[3] Vaidya, J., Clifton, C. W. & Zhu, Y. M. in Privacy Preserving Data Mining, pp. 1–5. (2006)

- Information on the legal context for TDM is current as of the date of publication of this report; the rapidly-evolving regulatory context may mean some elements become quickly out-of-date.
- During the period covered by this study the people of the United Kingdom voted to leave the European Union. The departure of the UK from the EU ('Brexit') may have significant implications for copyright law, research funding and international collaboration, among many other factors. However, it was not within the scope of this report to consider these issues.

## 1.6 Acknowledgements and declaration of interests

We would like to thank the ADBU for their invaluable input and involvement in the preparation of this report, and to Michael Jubb for his advice and guidance throughout the project. Our thanks also go to the wide range of experts and other stakeholders in the UK, France and elsewhere who kindly participated in our consultation and provided feedback on early versions of this report. A full list of contributors can be found in Appendix A.

Research Consulting is a provider of consultancy services to the Copyright Clearance Centre, Inc., whose RightFind for XML service is featured in case study 7. There are no other potential conflicts of interest relevant to our work on this report.

# Part A: TDM in context

# 2 The potential of TDM

*TDM is the automated analysis of a large quantity of digital content, which generates new information. By identifying patterns and correlations from unstructured information, it can provide the basis for discoveries and innovation, support competitiveness and contribute to national economic growth.*

## 2.1 Defining text, data and content mining

In today's digital world, the amount of information is increasing exponentially. More data has been generated by people and machines over the last two years than in the whole history of mankind[4]. Examples include customer information, user-generated content in social media, data from sensors and scientific instruments and an estimated 2.4 million scientific articles published every year[5]. This is so-called 'big data', a large portion of which is made up of natural language text.

Text and data mining (TDM) locates information within this mass of unstructured data. Just like conventional web searches, it applies computer algorithms to process a natural language text. However, TDM employs advanced software that not only compiles and analyses large amounts of digital information, but also makes connections that were not apparent before. TDM therefore has enormous potential to derive value effectively and efficiently from big data and the ever growing body of texts[6].

In its draft proposal for a Digital Single Market Directive, the European Commission defines text and data mining as:

*'Any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations'[7].*

For the purposes of this report, we distinguish three forms of TDM:

- **Text mining** analyses textual data as well as all other forms of data converted to text (e.g. audio transcripts). Its sources include:
  - books
  - newspaper articles

[4] IBM big data and information management; The Hague declaration. Big Data can reshape the world and save lives (infographic). (2016)

[5] Plume, A. & Weijen, D. van. Publish or perish? The rise of the fractional author. Research Trends. (2014)

[6] Hargreaves, I. et al. Standardisation in the area of innovation and technological development, notably in the field of Text and Data Mining. Report from the Expert Group. (2014)

[7] European Commission. Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market

  - web pages
  - social media posts
  - a wide range of other natural language resources
- **Data mining** targets all forms through which information can be transmitted: sounds, videos, images, graphs, numbers, chemical compounds, likes, clicks and others. Potential sources of this data include:
  - archives
  - databases
  - graphs
  - maps
  - sequences
  - formulas.
- **Content mining** is automated searching, indexing and analysis of the digital scholarly literature by software[8]. A distinct sub-specialism, it incorporates elements of both text and data mining in order to capture the diverse content contained within the scholarly record (natural language text, maps, formulae, graphs, tables, images, metadata and more).

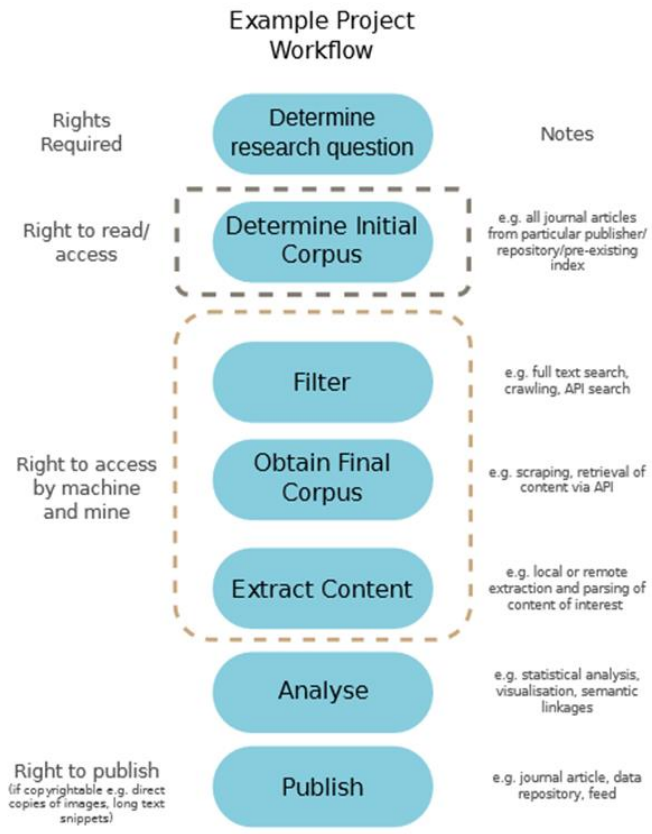
All three forms of TDM are used by researchers, and were referenced by stakeholders in this study. However, the primary focus of this report is on content mining of the scholarly literature.

## 2.2 The TDM process

In simple terms, the TDM process generally consists of the following steps: discover, download, normalise content and extract facts. Figure 1, reproduced from an article by Haeussler et al, illustrates a basic TDM workflow[8].

In practice, it is more complex. To start mining, a researcher needs to assemble a corpus of material. In principle, the research question, and, by extension, the research discipline would determine the choice of content. However, the initial corpus may not be readily identifiable as a discrete, well-defined entity. This is often the case in interdisciplinary research, where the researcher is a specialist in one domain, but less well-acquainted with the literature of other relevant domains.

**Figure 1 Example workflow for a TDM project**




[8] Haeussler, M., Molloy, J., Murray-Rust, P. & Oppenheim, C. Responsible Content Mining. (2015)

The researcher then needs to apply software to extract useful elements from the corpus, organising them and tracking useful relationships between them. This stage may involve significant time and effort, for example in the development of vocabularies.

Finally, the results can be analysed and shared. In most cases this would be done through writing up and publishing a paper. Increasingly, the underlying data will also be deposited for re-use by others, and linked to as 'supplementary data' from the associated articles.

## 2.3 Evaluating TDM's potential

By generating new insights and identifying patterns from largely unstructured information, TDM can provide the basis for discoveries and innovation, improve service accuracy and efficiency, support competitiveness and contribute to economic growth[9]. For instance, TDM has been used by businesses to analyse customer sentiment based on Twitter posts, to predict stock market fluctuations based on news feeds, to anticipate customer behaviour, and to tailor advertising strategies to specific customer segments. It has also been used by public authorities, for instance in crime prevention, intelligence and counter-terrorism.

> 'Text-mining has so many application domains, it is absolutely incredible'
> Petr Knoth, Open University

This report considers the use of TDM to generate new knowledge by extrapolating trends and information from scientific publications. Analysing scientific literature at scale facilitates the investigation of grand challenges such as climate change, disease treatment and prediction of epidemics. Mining of biomedical literature has already resulted in the discovery of new connections between chemicals and diseases – such as magnesium deficiency and migraine[10]. Using TDM to cross-examine neurological and life-style information can accelerate research into Alzheimer's disease[11]. More generally, text mining of scholarly papers can help track the evolution of ideas, examine logical connections between facts and findings or assess the impact of particular assertions[12].

Evidence from a range of sources suggests that the realisation of TDM's potential has been inhibited by the difficulties and costs of securing appropriate permissions to mine content[13]. Legislative changes could therefore deliver short-term benefits in the form of a reduction in the transaction costs associated with TDM (see next page).

[9] Einav, L. & Levin, J. The Data Revolution and Economic Analysis. Innovation Policy and the Economy. 14, 1–24 (2014)

[10] Baker, N. C. & Hemminger, B. M. Mining connections between chemicals, proteins, and diseases extracted from Medline annotations. Journal of Biomedical Information. 43, 510–519 (2010); Swanson, D. R. Migraine and magnesium: eleven neglected connections. Perspectives in Biology and Med. 31, 526–557 (1988)

[11] Greco, I. et al. Alzheimer's disease biomarker discovery using in silico literature mining and clinical validation. J. Transl. Med. 10, 1–10 (2012)

[12] Jisc. The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. 1–63 (2012)

[13] Boulanger et al. Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU. (2014) ; Jisc. The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. 1–63 (2012) ; IPO. Impact Assessment. (2012)

## THE ROLE OF A COPYRIGHT EXCEPTION IN REDUCING TRANSACTION COSTS

A 2014 analysis of policy options for the European Commission concluded that 'licensing in the market for TDM is potentially associated with (relatively) high transaction costs' (Boulanger et al, see prior page). The implementation of an exception to copyright allows these transaction costs to be mitigated, benefiting public research by:

1. **Saving researcher time**: The Wellcome Trust has estimated that obtaining permission to mine all articles on malaria in the Europe PubMedCentral database would take a full-time researcher nearly **8 months**, at a cost of over **€20,000**. A copyright exception potentially eliminates the need to obtain permission in this way.

2. **Cutting the compliance burden**: University College London have estimated that, before the adoption of a TDM exception in the UK, the costs of checking TDM compliance could amount to over **€500,000** per year. Their estimate was calculated based on the average salary for a team of 10 extra staff, plus the academic time needed to ensure that researchers are compliant.[14]

However, while businesses are already profiting from TDM, the impact of TDM on the research process itself is likely to materialise only gradually. This is common to early-stage technology, whereby investment in infrastructure predates returns (economic or social) by several years, or even decades. Nevertheless, a number of examples are already available which illustrate the efficiency savings from using TDM for research purposes, as summarised on the following page.

In the longer-term, improvements in efficiency of research can be expected to translate into downstream benefits for the economy and society as a whole. In a study for the League of European Research Universities (LERU), Biggar Economics estimated that for each €1 in gross value added directly generated by the LERU Universities, there was a total contribution of almost €6 to the European economy[15]. Meanwhile, an analysis prepared for the UK government has estimated positive and significant social returns of around 20% for UK public research and development investments. Furthermore, far from displacing private investment, empirical evidence suggests that public research and development incentives in fact generate additional private research and development[16].

More efficient research, underpinned by TDM, could result in an even greater economic contribution from universities and public research institutions – but this depends in part on an enabling copyright regime. In the following section, we therefore examine the current and proposed legal environments in the EU, the UK and France in relation to TDM.

[14] The Wellcome Trust. Wellcome Trust submission to IPO consultation on copyright. (2012); European Commission. Impact assessment on the modernisation of EU copyright rules. Part 1/3 (2016)

[15] Biggar Economics. Economic Contribution of the LERU Universities. (2015)

[16] Frontier Economics, Rates of return to investment in science and innovation: A report prepared for the Department for Business, Innovation and Skills (2014)

## IMPROVING RESEARCH EFFICIENCY THROUGH TEXT AND DATA MINING

Examples of TDM's potential to improve research efficiency include:

**Increased coverage**: Meta-analyses of published research or unstructured data sources have long been used to validate existing findings and discover trends and patterns. Evidence from the field of systems biology suggests the use of text-mining can **increase coverage by a factor of 4**[17], making it possible to analyse a much higher number of unstructured data points in a more efficient fashion.

**Reduced manual work**: Text-mining techniques were used in a 2011 review of young people's access to tobacco by researchers at the UK's Institute of Education. They found that it allowed them to identify the expected number of relevant studies with only **25% of the usual manual work,** concluding that the method '**is highly promising and may save significant time and money**'[18]. This result was achieved using relatively simple TDM tools – even greater reductions can be expected from using advanced TDM tools oriented towards 'higher' levels of language analysis.

**Improved curation:** Text-mining can assist with the curation of relevant information contained within scholarly articles into a relational database. New information is being generated at such a rate that database curators are struggling to keep up. It has been estimated that if context-relevant assisted curation were employed in a database with 100,000 new entries per annum, it could result in annual cost savings of up to **€70,000** and a **50% gain in productivity**[Error! Bookmark not defined.].

**Automated information retrieval**: A research team at the University of Edinburgh in the UK are using TDM to keep healthcare researchers updated on developments in their domain, quickly identifying high-quality, relevant results from the millions of publications produced each year. TDM allows researchers to zero in on the **top 25%** of papers which are most relevant to any given search query – a process that would take **2-3 years** to complete manually, at a cost of up to **€100,000**[19].

**Accelerated drug discovery:** It takes many years **(~12**) and substantial financial investment (**~€1.3bn**) to develop a new drug, and undesirable effects (e.g., toxicity) may only be discovered in late phase trials. Computational methods can cut the time and cost of drug development by bringing knowledge from literature together with high-throughput data sets to identify both known and new relationships between genes, pathways, drugs, environmental contaminants and diseases[20].

Other benefits are less quantifiable, but no less important. Simplified data collection, increased data sets, new research directions, higher quality research, and innovation in products and services have all been cited in past reviews of this area[21].

[17] Jisc. The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. 1–63 (2012)

[18] Thomas, J. & O'Mara-Eves, A. Reconceptualising searching and screening: How new technologies might change the way that we identify studies. in 19th Cochran Colloquim (2011)

[19] FutureTDM. Improving Uptake of Text and Data Mining in the EU: TDM Spotlight - The University Perspective. (2016)

[20] Gonzalez et al, Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery (2015)

[21] Hargreaves, I. Digital Opportunity. A Review of Intellectual Property and Growth. (2011)

# 3 The legal environment for TDM

*Regulatory changes are crucial to prevent TDM researchers in Europe falling behind the rest of the world. A new EC directive is likely to see a copyright exception for TDM become law across Europe in time. The UK adopted an exception in 2014, and France has recently followed suit - but areas of uncertainty remain.*

## 3.1 TDM and copyright

Intellectual property legislation can play a crucial role in enabling the use of TDM. TDM requires copying, adaptation and/or digitisation of the original work, and governments must consider how to balance rights-holders' interests against the innovation potential of TDM[23].

Within the last decade, Asia has overtaken Europe as the world's leading centre for academic TDM research, as measured by number of publications. China is the global leader in patenting text- and data-mining procedures, and second only to the United States in the number of text-mining based publications. Iran, Turkey and India have shown some of the fastest growth in TDM use in recent years[24].

> 'The present use of TDM in Europe is significantly lower than in the US and Asia, most probably due to limitations imposed by the European legal framework'
>
> EC FutureTDM project[22]

In comparison with Europe, TDM researchers in the US benefit from a more flexible copyright law based on the so-called 'fair use doctrine', in conjunction with extensive licensing[25]. Where fair use is found, no additional permission is required[25] although TDM uses falling outside the fair use criteria require an appropriate licence. Other 'fair use' countries include Israel, the Republic of Korea, Singapore and Taiwan[26].

Across Europe, exceptions to copyright are increasingly being enacted or proposed to facilitate TDM. However, many believe regulatory developments have been moving more slowly than in other parts of the world[27].

---

[22] European Commission, CORDIS. FutureTDM

[23] Haeussler, M., Molloy, J., Murray-Rust, P. & Oppenheim, C. Responsible Content Mining. (2015)

[24] Filippov, S. & Hofheinz, P. Text and Data Mining for Research and Innovation. Interact. policy *Br.* 1–16 (2016)

[25] Cox, K. L. Text and Data Mining and Fair Use in the United States. (2015); Wasoff, L. F. Text Mining and Fair Use

[26] Filippov, S. & Hofheinz, P. Text and Data Mining for Research and Innovation. Interact. policy Br. 1–16 (2016)

[27] Kroes, N. Our single market is crying out for copyright reform. (2014); Reda, J. Report on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI)). (2015)

## 3.2 Fair use versus exceptions to copyright

In understanding the impact of differing legal regimes on researchers' ability to conduct TDM, it is helpful to further consider the differing approaches taken in the US and Europe. In recent years, US case law and related guidance has clarified that:

- Copying of books for, inter alia, purposes of TDM is allowed when it 'poses no harm to any existing or potential traditional market for the copyrighted work'[28].
- Copies of the text and images mined may be retained, subject to fair use analysis[29].
- Fair use further extends to displaying parts of a mined text and images for commercial use, as long as it does not 'provide a significant market substitute for the protected aspects of the originals'[30].

The fair use doctrine offers researchers an 'affirmative defence' to charges of copyright infringement, but text mining is not itself a 'use' susceptible to fair use analysis. Instead, the question is whether the particular purpose for which the reproduction of the copyrighted materials is done qualifies as fair use[31]. Legally, whether text mining 'is' fair use thus remains a question to be answered on a case-by-case basis. However, the position of the US Association of Research Libraries on the mining of research articles is clear: 'In almost all cases, performing TDM on accessible articles is a fair use'[29]. Under the doctrine of 'copyright exceptions' prevalent in Europe, many argue that copyrighted material cannot be mined without a licence, *unless* an exception applies. The burden of proof then falls on the researcher to prove that they are meeting the specific criteria listed in the exception to copyright.

## 3.3 Overview of EU, French and UK legal environments for TDM

The UK enacted a copyright exception for TDM in 2014, and France in September 2016. The European Commission's recent proposal for a Directive on copyright in the Digital Single Market also makes provision for a copyright exception for TDM. Further details on the relevant legislation are provided in Appendix B. Though broadly similar, the scope and application of the three exceptions differs:

- The **French exception** is the most narrowly defined of the three. It is limited in terms of the research purpose (any commercial purpose is excluded), the sector performing the research (only 'for the needs public research'), and the subject matter to be mined (any kind of text, but only data 'including or associated with scientific writings'). It explicitly restricts the right to conserve and communicate copyrighted material to designated organisations, and makes no provision for publishers to deploy technical protection measures. The arrangements for regulation and practical implementation of the exception are still to be determined, and researchers await reassurance on how they might gain remote access to content for mining purposes.
- The **UK exception** is similarly limited to non-commercial research, but places no restrictions on who can perform the research, or the subject matter to be mined. The implication is that

[28] Authors Guild v. Hathi Trust, 755 F.3d 87 (2d Cir. 2014), Court ruling

[29] Association of Research Libraries. Issue Brief: Fair Use in Text and Data Mining. (2015)

[30] Authors Guild v. Google, Inc., No. 13-4829 (2d Cir. 2015), Court ruling

[31] Wasoff, L. F. Text Mining and Fair Use

anyone who is entitled to mine the material can retain the results. Results can also be communicated to anyone, as long as they do not contain copyright material beyond what is acceptable under the pre-existing quotation exception. Provision is made in UK Intellectual Property Office guidance for publishers to use technical protection measures.

- The **proposed EC exception** is the most liberal of the three. It permits TDM for both commercial and non-commercial scientific research purposes, but limits the benefit of the exception to 'research organisations'. Like the UK exception, it makes no specific provision for conservation and communication of copyrighted material, and allows publishers to deploy 'measures to ensure the security and integrity of the networks and databases' – provided these do not 'go beyond what is necessary to achieve that objective'.

**Table 1. Legal environment for TDM - lawful access to material is assumed in each case**

| | US 'fair use' | EC proposal | French exception (Loi pour une République Numérique) | UK exception |
|---|---|---|---|---|
| What activities are covered? | All activities, subject to fair use analysis | Scientific research | Scientific research | Research |
| Is a commercial purpose explicitly ruled out? | No, but the purpose is a factor considered when determining 'fair use' | No, as long as the research is done by a 'research organisation' (as per art 2) | Yes | Yes |
| Who is allowed to mine copyrighted material? | Anyone, subject to fair use analysis | Non-profit research organisations<br><br>Research organisations with public interest mission recognised by Member States | Public research organisations only | Anyone |
| Who is authorised to conserve copyrighted material? | Anyone, subject to fair use analysis | Not specified | Bodies designated by decree | Not specified |
| Who is authorised to communicate copyrighted material? | Anyone, subject to fair use analysis | Not specified | Bodies designated by decree | The copyright owner or anyone authorised by the copyright owner |
| Are rights-holders able to limit use of TDM? | Yes, in some circumstances (e.g. where provided for in the licence agreement) | Measures to ensure the security and integrity of the networks and databases | Not specified | Technical protection measures that are 'reasonable' |
| What kind of text and data can be mined? | Any, subject to fair use analysis | Works or other subject matter | Any kind of text, and data included in or associated with scientific research | Any work |

# Part B: TDM in Practice

# 4 Achieving legal clarity

The introduction of a copyright exception is not sufficient to empower researchers to undertake TDM. Without clear guidance on the exception's scope, and support for its application in practice, researchers remain reluctant to take advantage of the new opportunities presented.

## 4.1 Impact of the copyright exception in the UK

A copyright exception can allow researchers to undertake TDM on materials to which they have lawful access, without the need for a licence. The experience of UK researchers suggests that the introduction of an exception undoubtedly makes TDM easier, but it does not dispel all uncertainties about what is allowed. In the UK, copyright exceptions can be used in a court of law in response to copyright infringement charges, therefore putting the onus of responsibility on the person doing or facilitating the copying to ensure it is lawful[32]. No extensive right to use copyrighted material is granted, and in the absence of case law, few researchers, or librarians, appear confident in their ability to act within the law. Practical guidance, such as that prepared by Jisc[33], is crucial to assisting researchers and research support staff in this area.

So far, continuing uncertainty has created resistance to using TDM in the research community. Our interviewees expressed particular concern over the definition of 'non-commercial research', with many worried about the implications of working with non-academic partners (see case study 1).

**CASE STUDY 1: CONTINUING UNCERTAINTY**

CORE (www.core.ac.uk) is a UK initiative which aims to aggregate open access research outputs worldwide and make them available to the public. Since aggregation involves copying, the project's founder, Petr Knoth, faced repeated questions over the legal implications of CORE's activities in its early years. He notes that 'one particularly difficult case probably took two months of my life to resolve, and all for no benefit'.

The introduction of the 2014 exception has made a 'massive difference' to CORE's work, according to Knoth, but has not resolved all the issues. 'The definition of commercial vs non-commercial use is creating uncertainty', he explains. 'I don't see it as a problem for the commercial sector - if companies want to do it they will do it - it is just a problem for the university sector, where we are not allowed to take risks.'

[32] Secker, J., Morrison, C., Stewart, N. & Horton, L. To boldly go… the librarian's role in text and data mining. CILIP Update Magazine. (2016)

[33] Jisc. The text and data mining copyright exception: benefits and implications for UK higher education. (2016)

UK researchers also expressed concerns that even royalties from book sales might be deemed as commercial. A further issue is the use of technical protection measures to enforce digital rights management (DRM), which can prevent the use of TDM. Where a conflict arises between the exception and DRM, researchers can seek the mediation of the Secretary of State, but few seem willing to invoke such a process in practice. Professional and representative bodies therefore have an important role to play in safeguarding researchers' interests in these areas (case study 2).

A handful of researchers in the UK are already seeing the benefits of the exception, with European researchers even considering relocating to the UK to take advantage of the legislation (case study 6). Nevertheless, many researchers remain confused about what activities fall within the scope of an exception, and this is undoubtedly inhibiting its uptake.

**CASE STUDY 2: ACTING IN RESEARCHERS' INTERESTS**

The UK Libraries and Archives Copyright Alliance (LACA) and Universities UK work together to collect evidence of problems encountered by UK researchers trying to use the TDM exception. In late 2015 LACA was alerted to the case of a UK academic trying to mine a publicly accessible website, which used Captcha technology that prevented him from downloading more than a few records at a time.

LACA therefore submitted a complaint to the UK Intellectual Property Office (IPO), ultimately receiving advice that it was not within scope of the exception. Commenting on this outcome, a group of UK academic librarians observed: 'It is important to keep fighting battles even if you think it's unlikely you will win. Without further applications to government to address this issue on a case-by-case basis there will be no possibility to change the status quo'[34].

## 4.2 Impact of the copyright exception in France

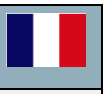
The situation in France under an exception to copyright will not be identical to the UK. French law makes no provision for publishers to operate technical protection measures, and the exception applies only to public research organisations. It is likely that French researchers will face challenges in determining whether research qualifies as non-commercial, and determining which works qualify as 'incluses ou associées aux écrits scientifiques'. Potential conflicts between the exception and the use of digital rights management (DRM) can also be anticipated.

The UK's approach has been to empower researchers to make use of the exception, with limited arrangements for governance or oversight. By contrast, French legislation identifies a specific role for named organisations to conserve and communicate technical copies, implying a greater degree of co-ordination and control. We note in particular the potential for the Bibliothèque nationale de France to play a role in this area (case study 3), and ISTEX's existing role in facilitating TDM of scientific articles under licence (case study 10). The Commission nationale de l'informatique et des libertés (CNIL)[35] and the Institute national de l'audiovisuel (INA) may also be able to help researchers navigate the legal implications of TDM in their areas of competence. Over-regulation of TDM, as in any other area, risks

[34] Secker, J., Morrison, C., Stewart, N. & Horton, L. To boldly go… the librarian's role in text and data mining. CILIP Update Magazine. (2016)
[35] Please see https://www.cnil.fr/ for more information

stifling innovation. Yet evidence from the UK suggests researchers require greater clarity and reassurance on the legal implications before they will adopt it at scale. Identifying appropriate mechanisms and organisations to deliver this clarity is a critical step to wider adoption of TDM.

## CASE STUDY 3: THE ROLE OF THE NATIONAL LIBRARY



The Bibliothèque nationale de France (BnF) has longstanding experience of balancing the needs of researchers against those of rights-holders. It already supports researchers in mining its web archive, within strict parameters, and acts as a trusted third party in the production of printed materials for the disabled.

Emmanuelle Bermes, the BnF's Deputy Director for Services and Networks, sees potential for BnF and other stakeholders such as ISTEX to help validate researchers' requests for access to copyrighted content. She stresses the need for a process to determine 'Who is asking for the data, and for what purpose?'. The aim would be 'to reassure the publishers that nothing bad will happen to the data, and to make sure that the people who are undertaking TM are allowed to do it'.

BnF could also play a role in preserving extracted corpora and research data for future re-use and replication. 'If we can have a lightweight process, with minimal infrastructure, that makes sure that the spirit of the law is honoured, that would be a good outcome for now,' she concludes.

## CASE STUDY 4: MINING OF SCIENTIFIC LITERATURE

Our consultation indicates that very few of the researchers employing TDM techniques in France base their work on scientific literature. In some cases this reflects the nature of the research question, or access issues, but a lack of enabling legislation is undoubtedly a factor. Very few academics are prepared to take the risk that data collection could be obstructed by copyright.

**Avoiding the copyright issue**

Julien Velcin, Associate Professor of Computer Science at the University of Lyon 2, relies on sources that are freely available on the Internet: 'I work mostly on abstract and title, not on the full publication, because this information is easy to gather from the web'. In a study of the evolution of the domain of geography, he gravitates towards material available to researchers from the social sciences and humanities web portal Persée[36]: 'We don't want to deal with copyrighted material other than within the Persée project…. we have very strong copyright issues in France, so for us it's very difficult to get this kind of information'.

**Change on the horizon**

With a TDM exception now in place, new possibilities will emerge for the mining of scientific literature. Amedeo Napoli, Head of the Orpailleur team at LORIA research laboratory in France observes: 'I think if we have easier access to full papers, for example, it will help us to develop other practices and open doors to different activities that can allow us to have better results'. Unlike in the UK, however, researchers such as Velcin who also mine social media sources are less likely to benefit, given the French exception's restriction to scientific texts. For Velcin, this means too many grey areas still remain: 'Data from sources like Twitter is somehow public data - but not completely public. There was no clear ruling on that in French law'.

[36] Please see http://www.persee.fr/ for more information

# 5 Delivering access to content

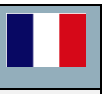
Researchers in the UK and France find their research constrained through inability to access content. Yet publishers receive relatively few requests for access, and argue they are already investing in services to deliver it effectively. Reconciling these competing narratives and interests is key to improving uptake of TDM.

## 5.1 Barriers to access

A copyright exception reduces legislative and contractual barriers to non-commercial TDM. However, it does not address the practical challenge of accessing the content to be mined. Researchers in both the UK and France cite difficulties in securing access to published content, and it is not uncommon for them to scale back their research as a result (see case studies 4 and 5). Content can be sourced through bulk downloads or web crawling, but in each case researchers encounter barriers to access:

- **Technical protection measures (TPMs).** Publisher may impose reasonable limits on speed or volume of downloads to ensure that mining does not corrupt or reduce quality of service.
- **Crawler traps.** Publishers also introduce so called 'crawler traps' - web pages that look like an academic paper to an automated downloading programme, and block access to the rest of the publisher's content where mining software is detected[37].
- **Restricted access to application programming interfaces.** APIs can be used to acquire content in machine-readable form. However, not all publishers offer an API, some require additional agreements before access is granted, and varying data formats make aggregation difficult.

### CASE STUDY 5: REINING IN RESEARCH AMBITIONS



In 2012, the French National Institute of Agricultural Research (INRA) initiated a project to classify fish species by their potential for domestication. Having identified 60,000 relevant papers, they embarked on a lengthy process of seeking publisher permissions to mine them. Mathieu Andro, the INRA project manager, estimates that it was a week's work to secure the licences, but publishers' technical protection measures presented a much greater barrier. Even with a team of people and the aid of automated tools such as Endnote, the original goal of mining 60,000 articles soon proved an impossible task.

Frustrated, Andro turned his attention to the literature available in ISTEX (case study 10). 'It was the best solution for us - but you don't have all the journals, particularly the most recent ones', he explains. 'It meant we had to abandon the list of 60,000 articles, because we knew we wouldn't find everything in ISTEX, but it was the only viable option in the circumstances'.

[37] Rachel Becker. Publisher under fire for fake article webpages. (2016)

## 5.2 Working with scholarly publishers

### CASE STUDY 6: 'PUBLISHERS STOPPED ME DOING MY RESEARCH'

Chris Hartgerink, a PhD student at Tilburg University in the Netherlands, uses tools developed by Contentmine in the UK (case study 11) to search Psychology articles for evidence of fabrication. However, gaining access to the articles he needs has proved far from simple. Finding much of his work blocked by publishers, Hartgerink says: 'I scaled down my TDM research, and had to exclude two publishers, meaning I was able to do some of the research, but not what I set out to do'.

Hartgerink is now exploring how to undertake his research with the Contentmine team in Cambridge: 'I know I can go to the UK and download these papers … because I am then geographically relocated I am allowed to do it.' Hartgerink sees the UK as a 'safe haven' for this type of research, but it is only a partial solution. He is still required to interact with many different publishers, and securing access to content remains a daunting task. In the longer term, he wants more publishers to offer freely accessible APIs, and to make better use of CrossREF's TDM service: 'For those publishers that don't apply additional conditions it works pretty well'.

Some scholarly publishers have already taken steps to facilitate text-mining of their content, both individually and collectively. In the UK, the Publishers Licensing Society (PLS) has established PLSclear TDM[38], a web service developed to reconcile the needs of publishers and researchers, and make it easier for publishers to manage requests for TDM. Other services are provided by CrossRef[39] and the Copyright Clearance Center, Inc. (case study 7).

Publishers stress that making data available in the form that researchers require is not always straightforward. Some publishing platforms are externally hosted, with the relevant technical protection measures being put in place by third party suppliers rather than publishers themselves. It can also be hard for publishers to distinguish legitimate downloads of material for TDM purposes from large-scale pirating of material. However, some have called into question claims that TDM represents a genuine threat to the stability of publishers' platforms[40].

Meanwhile, the Journal Article Tag Suite (JATS) is increasingly used to regularise and control content, defining a set of XML elements and attributes for tagging journal articles[41]. This helps to facilitate text-mining by allowing article elements to be identified consistently. However, as the standards evolve over time, older content may not be held in the same form as more recent publications. The effort required to update older publications is potentially substantial, and so variations in tagging practice may be found even within an individual publisher's archive. Publishers also note that demand for access to content since the introduction of the UK's exception has been muted. A 2015 survey by the PLS found no evidence of a significant increase in requests following the exception's introduction, with only 15% of publishers receiving requests. The majority were also granted within two weeks[42].

[38] Publishers Licensing Society. PLSclear TDM

[39] Crossref. Text and Data Mining Services. (2016)

[40] For example, Cameron Neylon has stated that PLoS servers are hardly affected by TDM downloads (2013)

[41] National Centre for Biotechnology Information. Journal Article Tag Suite.

[42] Publishers Licensing Society. Survey shows text and data mining supported by licensing not copyright exceptions. (2015)

### CASE STUDY 7: SOLVING THE MANY-TO-MANY PROBLEM

A challenge for both researchers and publishers is the sheer number of interactions required to obtain copyrighted materials at scale. The potential for a trusted third party to facilitate TDM on behalf of publishers and end-users has therefore long been recognised.

**The value of aggregation**

Copyright Clearance Center, Inc. developed its RightFind for XML service to address this need in the corporate world. 'Having a right to mine and being able to mine effectively are two different things' explains Roy Kaufman, CCC's Managing Director, New Ventures. 'People are asking whether they have a right to mine, but the question should be "how I can do it?"'

This is where a trusted third party can help, argues Kaufman: 'It's about not having to go to every publisher, not having to scrape, not having to get a feed. Everything takes time, so aggregation is really important'.

**Standardisation and security**

Kaufman also stresses the need for standardised data formats, and to respect publishers' concerns over data protection and security: 'It's hugely important to publishers that [their data] is not released into the wild'. Both publishers and content users are likely to ask searching questions about the credibility of any intermediary, the contractual terms applied, and their ability to deliver a viable service. All this is surmountable, but it is not something to be embarked upon lightly, concludes Kaufman: 'There's a lot of work - we've done it, so we know'.

## 5.3 Breaking the impasse

Researchers welcome publisher efforts to support TDM but feel there are still too many barriers to mining subscription-only content (case study 6). As a result, many prefer to mine open access sources such as Europe PubMed Central[43] or CORE (case study 1) or rely on services such as ISTEX (case study 10). This results in the low level of demand experienced by individual publishers, arguably undermining researchers' case that improved access is required.

Meanwhile representatives of the UK library community have stressed it is too early to draw conclusions on uptake, and note that much remains to be done to advance the TDM research agenda[44]. They also note that the volume of requests to publishers is not representative of the level of TDM occurring, since an exception allows researchers to mine content to which they have lawful access without seeking additional permissions.

It is clear that ongoing dialogue is needed between the research and publishing communities to find workable solutions to the access issue. Consortial bodies and publisher trade associations will have an important role to play in setting expectations on both sides (case study 8).

[43] An archive of life science journal literature

[44] Secker, J., Morrison, C., Stewart, N. & Horton, L. To boldly go… the librarian's role in text and data mining. CILIP Update Magazine. (2016)

## CASE STUDY 8: A CONTINUED ROLE FOR LICENCING

The adoption of a copyright exception is intended to reduce transaction costs, by eliminating the need to negotiate access on a case-by-case basis. In practice, publishers' rights to use technical protection measures under UK law means that disputes over access terms can persist even under an exception.

**Working with publishers**

Patricia Killiard is Acting Deputy Director, Academic Services at Cambridge University Library, and oversees licensing of scholarly content on behalf of the University. 'We've had some [licensing] examples that have been really urgent around full-text humanities content' she recalls. She cites the case of a US-based publisher, where researchers needed to secure access to licensed content as part of a research grant: 'The publisher sent them a form to sign, which basically meant trying to tie down what they could do, in complete opposition to the TDM exception'. The proposed terms meant that only single copies of the content could be made, and each individual project and researcher would need their own licence to obtain a copy. While the university stressed that researchers were entitled to mine the content as a result of the exception, the publisher argued that they were entitled to take 'reasonable steps to protect their content' - even though the UK TDM exception explicitly excludes the possibility that it can be overridden by contract.

**Finding a way forward**

While this was an extreme example, it illustrates the continued difficulties faced by researchers seeking to mine content under the copyright exception. Killiard believes consortial bodies such as Jisc Collections in the UK and the Couperin Consortium in France have an important role to play in breaking the impasse over access. These bodies work with publisher trade associations to agree model licence terms, and the adoption of clauses covering TDM could support the practical implementation of a copyright exception. 'Licences allow you to do things that you can't do under the law', Killiard explains. 'The law permits publishers to apply technical protection measures, so there has to be some negotiation about what publishers need to do to protect their content'. Nevertheless, previous efforts in this area, such as the EC-convened "Licences for Europe" initiative, encountered difficulties finding common ground between researchers, librarians and publishers[45]. Ultimately, then, the library community may need to be prepared to test matters in court. So far, says Killiard, 'We haven't been brave enough, we haven't really tested these exceptions legally'.

[45] Van Noorden, Tensions grow as data-mining discussions fall apart, Nature. (2013)

# 6 Developing infrastructure

Text-mining at scale cannot take place without infrastructure. Investment is needed in the technologies used to aggregate, normalise, interrogate and preserve TDM materials. In the medium-term, the goal should be to develop user-friendly cloud services, which can be used by researchers with minimal technical expertise.

## 6.1 The importance of infrastructure

Mining of scientific publications at scale relies on a complex infrastructure of text repositories, archives, databases, aggregators, text and data mining tools, data storage services and publishers' APIs for accessing scientific publications and research data. Without a well-developed infrastructure, researchers must painstakingly download scientific articles from individual publishers, and assemble a corpus of material manually. This is a very inefficient process, yet it is the approach taken by many of the interviewees consulted in our work. Achieving major gains in research competitiveness and efficiency requires much greater development and adoption of shared infrastructure and software tools. However, there is little certainty that such development will occur.

The non-commercial exceptions in the UK and France reduce the prospect of commercial returns from TDM for researchers and research organisations. Their existence also discourages publishers and content providers from investing in infrastructure[46]. In order to deliver the desired increase in efficiency, investments must therefore be made from public sources. At present, there appears to be a greater willingness to invest in TDM resources and tools for public research in France than the UK.

## 6.2 TDM infrastructure in the UK and France

Both UK and French researchers are participants in the EC's OpenMinTeD project. This aims to develop an interoperability framework to allow text mining research communities and service providers to deliver and consume text mining tools in a seamless and uniform way[47]. The project seeks to address a significant gap in current TDM infrastructure, but must contend with the fragmented copyright regime at European level.

At a national level, the UK moved rapidly to adopt a legal exception, but has made only limited investment in the technical and human infrastructure needed to support TDM. There are longstanding centres of excellence in Manchester and Sheffield (see case study 9), as well as Edinburgh and Cambridge. Nevertheless, support for TDM lacks critical mass, and the majority of researchers rely on open-source software and publicly-available content. The UK's commitment to gold open access may

[46] Boulanger et al. Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU. (2014)
[47] See http://openminted.eu/ for more information

prove advantageous for TDM in the long-term, by opening up more content for mining. However, at the present time the majority of scientific literature remains behind a paywall, and only ContentMine are making a concerted effort to develop the tools, standards and infrastructure needed to access it (case study 11).

By contrast, France's investment in ISTEX (case study 10) has created a critical mass of content, infrastructure and expertise that bodes well for the future. Researchers such as Mathieu Andro at INRA (case study 5) are using commercial software for text-mining purposes, while others have been able to secure limited funding to access proprietary databases (case study 14). Nevertheless, only a handful of the French researchers we consulted are choosing to text-mine the scholarly literature. Mining of databases, web content and social media remains far more common. French researchers face additional challenges when seeking to mine content in their native language. TDM tools can be language-independent or dependent, but in the case of the latter, work done by the EC's FutureTDM project[48] finds, unsurprisingly, that English remains the language of almost all software packages[49]. French is part of a second tier of moderately well-supported European languages[50].

## CASE STUDY 9: SOFTWARE AND TOOLS FOR TEXT-MINING

GATE (www.gate.ac.uk) is a large framework of open source software for text mining, whose aim is to develop tools and methods to make text mining as easy as possible. Based at the University of Sheffield in the UK, the GATE team collaborate with universities, research labs and companies who have large collections of texts, but lack text mining capabilities to extract useful information from them. GATE's tools have been downloaded about 300,000 times since 2005.

Mark Greenwood, a member of the GATE team, explains its value as follows: 'Pretty much every time you have a new project or data source that you want to process you hit issues about how the documents are structured, oddities of formatting, and so on.' GATE's tools can convert this varied material into a standard form, creating a foundation for further processing and to answer specific research questions.

The UK's National Centre for Text Mining (www.nactem.ac.uk) at the University of Manchester has also recognised the need for easy access to text mining tools. NaCTeM has been developing a Web-based text mining workbench, Argo[51], which also supports collaborative work, e.g., by curators of scientific databases. This allows users to configure workflows consisting of text mining components and has been deployed in cloud and cluster environments for processing of large-scale collections. Argo workflows have also been used to process the content of Europe PubMed Central (>3 million full text documents), the archive of the British Medical Journal (380,000 articles since 1840) and the content of the Biodiversity Heritage Library (50 million pages).

[48] European Commission. FutureTDM. The Future of Text and Data Mining. (2016)

[49] Pouli, K. Techniques, Tools and Technologies for TDM. (2016)

[50] META-NET White Paper Series: Key Results and Cross-Language Comparison (n.d)

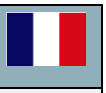
[51] See http://argo.nactem.ac.uk/ and Rak, R., Rowley, A., Black, W.J. and Ananiadou, S. Argo: an integrative, interactive, text mining-based workbench supporting curation. Database: The Journal of Biological Databases and Curation. (2012)

## 6.3 Looking to the future

There is great potential for cloud services to support text-mining in the future. One interviewee suggested this might take a form similar to Google's app engine[52], which provides a platform for building scalable web applications. Others foresee cloud services which require little or no technical expertise, and can be run on corpuses hosted remotely – whether by publishers or services such as ISTEX. This is also one of the goals of the OpenMinTeD project on TDM infrastructure and interoperability[53]. One researcher concluded: 'I am fully convinced that when TDM becomes more accepted and developed there will be programmes as easy to use as SPSS[54]. Once that happens then I think the uptake will rise greatly.'

Making the most of these opportunities will require sustained investment from multiple sources. Investments in TDM infrastructure are needed to increase the competitiveness and efficiency of scientific research, with spill-over effects on innovation and economic growth. TDM poses challenges that cannot simply be solved through legislation: aside from removing legal barriers, policy makers must also consider ways to drive investment in infrastructure to fully realise TDM's potential.

### CASE STUDY 10: ISTEX – AN INVESTMENT IN THE FUTURE



The ISTEX project is a vast programme for the acquisition of scientific resources, aimed at setting up a digital library for the benefit of members of higher education and research establishments across France. With over €55 million invested in content acquisition since 2012, a further €5 million has been spent on hardware and software infrastructure.

Data from publishers is grouped into a single, normalised corpus, allowing researchers to mine thematic sub-corpuses, regardless of their origin. Researchers can also request the right to download sub-corpuses from ISTEX and use their own tools to mine the material. Any member of a French higher education or research establishment is able to benefit from ISTEX's wide range of software tools and powerful search engine.

The project budget also supports a dedicated team responsible for pre-processing, standardisation and enrichment of data and metadata. Meanwhile, the preservation of the data held within ISTEX is assured by the Institute for Scientific and Technical Information (INIST). In many respects, the service is a text miner's dream. Where their research questions can be answered within the confines of the ISTEX corpus, it offers French researchers a huge advantage over their counterparts in the UK and elsewhere. Where the material to be mined is of recent origin, or cannot be found within ISTEX, researchers in both countries face common challenges.

[52] See https://cloud.google.com/appengine/ for more information

[53] In OpenMinTeD, the Greek Research and Technology Network (GRNET) is leading the infrastructure provisioning activity, offering cloud computing resources to the project.

[54] A software package widely used by researchers for statistical analysis

# 7 Enhancing skills and support

TDM requires a high level of digital literacy. Text-mining experts, computer science departments and libraries can play a role in supporting researchers in acquiring the necessary skills. Improved levels of collaboration between text-mining experts and domain scientists will be necessary as TDM evolves.

## 7.1 Addressing the skills gap

There is a mismatch between the potential applications for TDM in public research and the existing skills base. On the one hand, researchers wishing to use TDM themselves need a high level of digital literacy. For instance, adjusting software parameters to suit a given research project generally involves programming – and such adjustments may be needed at different stages of the mining process. On the other hand, technical experts are often unaware of TDM's potential as an academic research tool. To address this problem, training and support can be provided to researchers by TDM communities and computer science departments at universities. Libraries can also build on their experience in information skills and provide training and support for TDM. Alternatively, research groups may choose to recruit experts with text-mining expertise to work on relevant projects.

### CASE STUDY 11: CONTENTMINE

ContentMine is a not-for-profit company in Cambridge, UK, whose declared mission is 'to liberate scientific facts from academic journals' and enable anyone to do research using TDM. Started in 2014, its founder Peter Murray-Rust received funding of some €300,000 from the Shuttleworth Foundation to demonstrate the usefulness of content mining and develop tools and services.

ContentMine (www.contentmine.org) provides guidance on how to get papers, create a normalised corpus, and process them to search for key terms. The team are often working in uncharted waters. 'There are no TDM standards. ContentMine is pioneering all of this', said Murray-Rust, a committed advocate of TDM, 'I hope to show that knowledge should be for everybody, not just for very rich universities who use tax payers' money to pay publishers'.

The company runs workshops to promote the value of TDM to researchers grappling with overwhelming levels of content. Talks by the ContentMine team have now reached an estimated audience of 2,000 people. Most recently, six young researchers from around the world, including one based in France, have been appointed as the first cohort of ContentMine Fellows – early adopters who want to use TDM to fast-forward their research.

'Two years [since the UK exception was introduced] is very little,' observes Murray-Rust. Yet he also stresses that an exception alone is not enough, and that there has been insufficient resource put into skills, software and legal support across Europe. Unless more is done to address these gaps, he warns, 'Europe will become a third-world nation in text and data mining'.

### CASE STUDY 12: TRAINING THE NEXT GENERATION OF RESEARCHERS

François Rioult at the Université de Caen is responsible for equipping students and researchers with the skills needed to make effective use of TDM. One point he emphasizes is that 'the data scientist has to be present from the very beginning of the process to the very end, and to master every part of every stage.' It's a hard job, he concludes, and many data scientists struggle to present their results effectively to a domain specialist. Conversely, researchers without expertise in TDM often fail to see the opportunities it presents. 'The world is now waking up to the fact we have lots of data', notes Rioult, 'but people continue to ask [research] questions from the previous century'.

## 7.2 The role of the library

Academic libraries have an important role to play in supporting researchers to use TDM. Following adoption of the UK exception, the library community has shifted its focus from lobbying for legislative change to encouraging its use in practice. Recent advice from UK librarians stresses the need for librarians to work in partnership with researchers[55], and develop TDM support as a service (case study 13). International publishers will often be unaware of national legislation such as a copyright exception, and may seek to apply standard contractual terms that conflict with the exception. Librarians therefore have a responsibility to gain an understanding of the new legal framework, and be prepared to challenge publishers and other content providers where appropriate. They must also be in a position to escalate issues to the appropriate authorities where necessary (case study 2).

### CASE STUDY 13: ENABLING TDM AT THE UNIVERSITY OF CAMBRIDGE

Uptake of the copyright exception by University of Cambridge researchers has been limited, according to Dr Danny Kingsley, Head of the University's Office of Scholarly Communication. While initiatives like ContentMine (case study 11) are leading the way, the majority of researchers are either unaware of the exception's existence, or unclear how it could benefit their research. Kingsley cautioned that 'this doesn't mean it's not happening at all but the research community is not thinking "we need to go to the library to talk about this"'.

The library has begun putting in place mechanisms to handling TDM licensing queries more effectively, and clarify internal responsibilities. A longer term goal is to equip librarians with the legal and technical expertise to support researchers effectively: 'The kinds of skills and knowledge that are needed now - it's a whole new world,' observes Kingsley.

Libraries will also need to be proactive in managing risk to the institution, representing researchers' in negotiating access to content, and developing an institutional position on the use of TDM. Much of this comes down to providing guidance to researchers on what TDM is, and identifying named people in the library who can support them. The aim, according to Kingsley, is to get to the point where the library can say, 'If you've got a question about TDM, come to us'.

[55] Secker, J., Morrison, C., Stewart, N. & Horton, L. To boldly go… the librarian's role in text and data mining. CILIP Update Magazine (2016); Stewart, N., Secker, J., Morrison, C. & Horton, L. Impact of Social Sciences – Liberating Data: How libraries and librarians can help researchers with text and data mining. LSE Impact Blog. (2016)

# 8 Providing incentives for TDM

TDM faces many of the same challenges as the wider move towards open science. TDM Researchers are often pioneers, and find it difficult to secure scholarly recognition and rewards for work of this kind. Funders and policy makers will need to develop greater incentives for TDM research before it can truly flourish.

## 8.1 TDM and open science

The promotion of TDM should be set firmly in the context of moves across the world towards open science. The 2002 Budapest Open Access Initiative makes clear that the ability to crawl articles for indexing and pass them as data to software is a central goal of the open access movement[56]. Developing the infrastructure and the broader environment to support TDM is a key step in moving from open access for publications towards fully open science. The link between TDM and open science is evident in the way TDM builds on open access literature, and in the aggregations of content provided by repositories (see case study 1).

As with open science, there are many drivers for the adoption of TDM, including:

- the availability of new technologies and services,
- inexorable increases in the volumes of content,
- the growth of research collaborations and international partnerships, and
- the policies of Governments and funders.

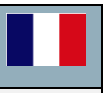
However, open science and TDM also face similar challenges. These include:

- simple lack of awareness;
- the strength and variety of disciplinary cultures;
- the investment of effort needed to gain the necessary skills and conduct the first experiments;
- the lack of widely-available and easy-to-use infrastructure and tools;
- doubts as to the likely value of the results;
- limited funding for data-sharing and curation; and
- concerns about copyright and licensing, and their impact on data-sharing (case study 14).

The incentives for researchers to use TDM themselves, train others in its use and develop TDM infrastructure are weak, especially given the pressure to publish in high-status scholarly journals. There are few examples of high-profile researchers securing reputation and rewards through their use of TDM, or of calls from major funders for proposals that make use of TDM techniques.

[56] Budapest Open Access Initiative (2002)

### CASE STUDY 14: MEETING THE COST OF CONTENT



David Chavalarias, Director of Paris's Institute des Systèmes Complexes, set out 10 years ago with an ambitious goal: 'to build a picture of science and its evolution over time'. Delivering this goal requires access to mine interdisciplinary digital archives, such as Thomson Reuters's Web of Science database. Chavalarias explains, 'We had to negotiate, we had to buy the databases… despite the fact that CNRS has a subscription'. With limited funding available, continued access relies on the good will of the provider, and can be cancelled anytime. The risk to long-term scientific projects is considerable, said Chavalarias: 'You just can't do sustainable TDM if you don't buy the database at full cost.'

Even where access is permitted, making the outcomes available to others is not straightforward. Chavalarias noted a number of areas of uncertainty: 'Can I process this data? Can I put it online for non-commercial use? Should I control who has access, or should it be open to the public?' Addressing these concerns and facilitating collaboration between researchers is critical to maximising the benefits of TDM.

## 8.2 Creating incentives

Addressing the lack of incentives for TDM will require the concerted action of stakeholders at institutional, funder and policy levels. Libraries and repository managers have key roles to play in supporting TDM, building on their close contacts with researchers, understanding of open access, knowledge of their collections, and understanding of copyright and licensing issues. However, policy-makers and funders at international, national and local levels must also develop financial and non-financial incentives to promote the wider adoption of TDM, and the benefits it can bring.

### CASE STUDY 15: CULTURAL BARRIERS TO TDM

A postdoctoral researcher at the University of Cambridge, Dr Ross Mounce has used text-mining across a range of projects in the fields of palaeontology and phylogenetics, both in academia and at London's Natural History Museum (NHM). Mounce's work benefited substantially from the UK's exception: 'As soon as the UK legislation came in that emboldened me, because I knew I could legitimately download a much greater quantity of material'.

For Mounce, the benefits of the exception cannot simply be measured in person-hours saved. 'It's really binary,' he explained, 'I could try to give you an estimate for how long it would take to [individually] download 10,000 papers, but that would be silly - I just wouldn't have done it'. He points instead to the benefits of his work, particularly in helping the Natural History Museum measure its impact on research.

While he remains a strong advocate of the UK's exception, Mounce is also clear on its limitations: 'the restriction to non-commercial use has been a huge hindrance to my work. If I was allowed to sell this as a service… I could develop a start-up around that. But that isn't going to happen in the UK'. He also points to a lack of training, skills and incentives, as well as cultural resistance to text-mining in many parts of academia. 'There are so many obstructions in the way of doing this research and doing it well, it is just too hard and so people do other things,' he concluded. 'I am in that position myself - there are no budgets out there, it is not even on the radar that this is possible'.

# 9 Conclusions

The use of TDM for public research remains limited in both the UK and France. The introduction of an exception provides benefits to existing TDM practitioners, but will not result in widespread uptake on its own. Changes to copyright law must be accompanied by improvements in access, infrastructure, skills and incentives for TDM to deliver its anticipated economic and societal benefits

## 9.1 The current state of TDM

This report provides case studies from a range of TDM researchers in both the UK and France, many of whom are associated with centres of excellence in text-mining. However, we encountered difficulties in identifying more than a handful of researchers who actively use, or wish to use, TDM on scholarly literature. Our work did not constitute a systematic survey, but it suggests that the use of TDM within public research remains relatively low, and that the technique is deployed within only a limited range of scientific disciplines at present.

The adoption of an exception to copyright within the UK has nevertheless delivered real benefits for a small number of researchers. In addition to the cases cited in this report, other instances have also been reported to the UK IPO, notably in the field of medicine and biology[57]. However, the benefits are not yet widespread, and the introduction of an exception has not stimulated rapid growth in uptake of TDM. These findings are consistent with a 2014 study of the likely impact of an EC exception, which anticipated only small increases in the rate of TDM for scientific purposes following an amendment to copyright law[58]. The slow rate of adoption can be attributed to the range of barriers to TDM outlined in our report, namely:

- Continued uncertainty over the exception's scope and application
- Difficulties in gaining access to content
- Inadequate infrastructure
- Gaps in skills and support
- Lack of funding and incentives

The case for a copyright exception for TDM has been made successfully in the UK, France and Europe on the basis of greater efficiency and reduced rights clearance costs for publicly-funded researchers.

[57] Stewart, N., Secker, J., Morrison, C. & Horton, L. Impact of Social Sciences – Liberating Data: How libraries and librarians can help researchers with text and data mining. LSE Impact Blog. (2016)

[58] Boulanger et al. Assessing the economic impacts of adapting certain limitations and exceptions to copyright and related rights in the EU. (2014)

Our work confirms that the legal and transactional barriers to TDM without an exception are real, but that technical and cultural factors are equally, if not more, important. The legal obstacles to TDM are being reduced, though more can be done to give researchers confidence in their rights. Attention must now turn to overcoming the remaining barriers to widespread use of TDM in public research, and we make the following recommendations accordingly.

## 9.2 Recommendations for libraries

There is much that libraries and library consortia can do to support and enable TDM. The key recommendations arising from our work are as follows:

1. **Develop mechanisms to monitor researchers' experiences** of using an exception, and be prepared to act on their behalf in raising legal issues with the relevant authorities.
2. **Develop case studies and guidance** on the types of research and organisations that fall within the scope of an exception, particularly with regard to commercial considerations. The guide prepared by Jisc in the UK provides a useful model for this purpose[59].
3. **Secure the support of the national library** in developing good practice, showcasing the possibilities of TDM, and developing a constructive dialogue with publishers.
4. **Incorporate TDM clauses into model licence agreements** which reference the exception and set out access arrangements for TDM. Consortia such as Jisc Collections (UK) and Couperin (France) have a key role to play in working with the publishing community to achieve this.
5. **Invest in dedicated TDM support services to assist researchers**, equipping library staff with the skills needed to support TDM and embedding them in TDM research groups (see box below).

### ENVISIONING A LIBRARY SUPPORT SERVICE FOR TDM

A fully-developed library support service for TDM would work in partnership with researchers to offer the following:

1) Advocacy for the benefits of TDM at all levels of the organisation
2) Copyright advice on using the TDM exception, any licence restrictions that can be ignored and how to attribute any sources, particularly if open data is used
3) Access to legal expertise
4) Skills development in indexing and metadata curation, and access to technical training in coding or the use of high performance computing (HPC) facilities.
5) Advice on data sources and tools available both in library collections and more widely online

[59] Jisc. The text and data mining copyright exception: benefits and implications for UK higher education. (2016)

## 9.3 Recommendations for other stakeholders

### Legislators

1. **Communicate legal provisions for TDM with certainty and clarity**, ensuring that the exception cannot be overridden by contract, and that regulatory arrangements are clear
2. **Clarify the exception's scope** in cases where public researchers collaborate with commercial partners
3. **Monitor the interaction of the copyright exception with digital rights management (DRM), licensing and other relevant legal regimes**, developing procedures to manage any conflicts which may arise.

### Research leaders and institutional managers

1. **Communicate the benefits of TDM** to the research community
2. **Invest in the development of library services to support TDM** including advocacy, legal advice, training and skills development
3. **Explore opportunities for knowledge exchange in TDM with commercial partners** including licensing and the development of spin-out companies

### Research funders and policy makers

1. **Invest in the infrastructure** needed to support TDM, developing a critical mass of content and technical resources for researchers.
2. **Identify or create shared fora** for representatives of researchers, librarians and publishers to discuss continued challenges over access, and develop solutions.
3. **Consider the needs of TDM researchers** as part of broader initiatives to promote open science.

### Publishers and infrastructure providers

1. **Pursue the development of cloud services for TDM**, with the goal of making TDM accessible to researchers without specialist technical expertise.
2. **Develop arrangements to streamline access to all forms of copyrighted material for mining purposes**. In France, this will require close co-operation between publishers, ISTEX, the BnF and INA.
3. **Adopt open standards and harmonised data formats** to facilitate mining of richer content from multiple sources.

# Appendix A – Interviewees and contributors

The following individuals were interviewed in the course of our work.

| Name | Role | Organisation |
|---|---|---|
| Mathieu Andro | Digitization and text mining projects manager | Institut national de la recherche agronomique (INRA), France |
| Elton Barker | Reader in Classical Studies | The Open University, UK |
| Emmanuelle Bermes | Assistant to Director of services and networks | Bibliothèque nationale de France (BnF) |
| David Chavalarias | Researcher at the CNRS and Director of the Institute for Complex Systems | Centre national de la recherche scientifique (CNRS), Institut des Systèmes Complexes, France |
| Stephanie Dales | Economic Advisor | Intellectual Property Office (IPO), UK |
| Mark Greenwood | Research Associate in the Natural Language Processing | University of Sheffield, GATE, UK |
| Chris Hartgerink | PhD student in Statistics | Tilburg University, the Netherlands |
| Margaret Haig | Head of Copyright Delivery | Intellectual Property Office (IPO), UK |
| Frank Hellwig | Business Development Manager | Ubiquity Press, UK |
| Laurence Horton | Data Librarian | London School of Economics and Political Science |
| Roy Kaufman | Managing Director | New Ventures, Copyright Clearance Center (CCC), US |
| Douglas Kell | Professor of Bioanalytical Sciences, formerly CEO of Biotechnology and Biological Sciences Research Council | University of Manchester, UK |

| | | |
|---|---|---|
| Patricia Killiard | Acting Deputy Director, Academic Services | University of Cambridge, UK |
| Danny Kingsley | Head of Scholarly Communication | University of Cambridge, UK |
| Petr Knoth | Research Fellow | The Open University, UK |
| Ross Mounce | Postdoctoral Researcher | University of Cambridge, UK |
| Peter Murray-Rust | Co-director, Professor Emeritus of Chemistry | ContentMine, University of Cambridge, UK |
| Amedeo Napoli | Senior Researcher CNRS, Head of the Orpailleur team at LORIA research laboratory | Laboratoire Lorrain de Recherche en Informatique et ses Applications, France |
| Charles Oppenheim | Independent Consultant, Professor Emeritus of Information Science | Chartered Institute of Library and Information Professionals, UK |
| Jean-Marie Pierrel | Professor of Computer Science | Initiative d'excellence en Information scientifique et technique, France |
| Thierry Poibeau | CNRS Director of Research and head of the LaTTiCE laboratory | Langues, Textes, Traitements informatiques et Cognition (LaTTiCE), France |
| François Rioult | Researcher | Groupe de recherche en informatique, image, automatique et instrumentation de Caen, Université de Caen Normandie, France |
| Jane Secker | Copyright and Digital Literacy Advisor | London School of Economics and Political Science |
| Julien Velcin | Associate professor of Computer Science | Université Lyon 2, France |
| Anna Vernon | Collections services manager | Jisc, UK |

We would also like to thank the following individuals who responded to requests for suggested interviewees or provided comments and feedback used in the preparation of this report: Sophia Anandiou (University of Manchester, UK), Rachel Bruce (Jisc, UK), Jérôme Darmont (Université de Lyon 2, France), Erik Ketzan (Birkbeck, University of London, UK), Alexandre Hannund Abdo (INRA, France), Angus Roberts (University of Sheffield, UK), and Frédéric Jurie (Université de Caen Normandie, France), Stuart Taylor and Helen Duriez (Royal Society, UK).

# Appendix B – The EU, French and UK exceptions to copyright for TDM

## 1. Europe

Use of TDM in Europe is currently regulated by EU Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society[60]. While the Directive does not currently permit creating copies of copyrighted materials for TDM, article 5 allows member states to introduce a number of discretionary exceptions and limitations (ibid.) in order to achieve public policy objectives such as education or access to information. However, use of article 5 has created a fragmented regulatory regime for TDM across the Union. This remains problematic despite efforts to harmonise and modernise EU copyright law[61] and advance TDM[62].

The Digital Single Market (DSM) Strategy adopted in May 2015 called for the EU copyright framework to address a set of key obstacles to the functioning of the DSM. Furthermore, the strategy announced legislation 'to reduce the differences between national copyright regimes and allow for wider online access to works by users across the EU', notably as regards cross-border access to copyright-protected content and exceptions in the areas of education and research[63]. Following the strategy, the European Commission carried out an impact assessment of EU copyright rules which included text and data mining among the activities worthy of legal exception as being relevant for access to knowledge, education and research[64]. The assessment recommends introducing an exception to copyright law which would be only applicable to 'public interest research organisations covering TDM for the purposes of both non-commercial and commercial scientific research'[65]. The exception is expected to increase legal certainty and reduce rights clearance costs for research organisations. This option was preferred to a more limited exception (covering TDM for non-commercial scientific research) and to a more liberal one (which would have allowed anyone who has lawful access to copyright material to use TDM for scientific research purposes of both non-commercial and commercial nature).

---

[60] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society

[61] Kroes, N. Our single market is crying out for copyright reform. (2014); Reda, J. Report on the implementation of Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (2014/2256(INI)). 2015; Reda, J. EU copyright evaluation report. (2016)

[62] European Commission, CORDIS. FutureTDM. Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach. (2015); European Commission. OpenMinTed; National Institute for Text and Data Mining (NacTeM)

[63] Jisc. The Value and Benefit of Text Mining to UK Further and Higher Education. Digital Infrastructure. 1–63 (2012)

[64] European Commission. Commission staff working document. Impact assessment on the modernisation of EU copyright rules. Part 1/3

[65] European Commission. Commission staff working document. Impact assessment on the modernisation of EU copyright rules. Part 2/3

The option suggested in the impact assessment was recently confirmed in a draft Commission proposal for a Directive on copyright in the Digital Single Market, but with some changes[66]. Article 3 of the proposed directive (superseding Directive 2001/29/EC article 5) asks Member States to provide for a copyright exception

> **'for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject matter to which they have lawful access for the purpose of scientific research' (ibid).**

Furthermore, article 2(1) defines research organisations as

> **'a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to conduct scientific research and provide educational services:**
>
> **(a) on a non-for-profit basis or by reinvesting all the profits in its scientific research; or**
>
> **(b) pursuant to a public interest mission recognised by a Member State;**
>
> **In such a way that the access to the results generated by the scientific research cannot be enjoyed on a preferential basis by an undertaking exercising a decisive influence upon such organisation' (ibid).**

The Commission's formulation therefore does not seem to rule out the possibility that TDM data be used in the context of a commercial research project, as long as such a project is undertaken by a public interest or non-profit research organisations. Moreover, the latter part of article 2(1) implies that all results obtained by the research organisation (including through TDM) have to be made public and cannot be enjoyed 'on a preferential basis'.[67] The proposal regulates which entities can enjoy the TDM copyright exception rather than focusing on the circumstances under which the exception can be granted.

Finally, the proposal states that 'rightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject-matter are hosted', but adds that 'such measures shall not go beyond what is necessary to achieve that objective'. It concludes by requiring Member States to encourage rights holders and research organisations to define best practices in this regard.

## 2. The UK exception to copyright regulations covering TDM

The UK was one of the first EU member states to introduce a series of exceptions to its copyright law, including a TDM exception. In 2011 the UK Government commissioned a review of Intellectual Property and Growth to address the concern that 'the current intellectual property framework might not be sufficiently well designed to promote innovation and growth in the UK economy'[68]. Among

[66] European Commission. Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market

[67] This would suggest, for instance, that research projects that are led by a university but co-financed by private company, and whose secondary objective is to deliver value to the partner organisation (e.g. through increased efficiency, improved products etc.), may still benefit from the exception

[68] Hargreaves, I. Digital Opportunity. A Reveiw of Intellectual Property and Growth

other concerns, the review recognised that copyright protection can unduly slow the dissemination of facts, data and ideas that are essential for innovation and economic growth. Subsequently, the UK Intellectual Property Office (IPO) conducted an assessment of the economic impact of removing copyright protection for certain public uses[69]. It estimated that reforms to copyright law could generate £250 million (€300 million) for the UK economy[70]. Introducing a copyright exception for TDM was predicted to reduce administrative costs associated with licensing while at the same time reducing delays to research; moreover, public authorities were expected to benefit from TDM research at a lower cost[69].

Following the IPO assessment, section 29a was introduced to the Copyright, Designs and Patents Act 1988 which removed barriers to TDM for non-commercial research purposes[71]. It established that

> **a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose**

In October 2014 the IPO clarified that lawful access is achieved if the researchers 'already have the right to read the work', in which case no additional permissions are needed[72]. In theory, TDM would thus be permissible where works are rented or on short-term loan, but arguably the corpus would have to be destroyed on expiry of the loan period. The exception also states that copyright holders are prevented from obstructing these new rights by using small print in contracts or terms of service and that 'contract terms that stop researchers making copies of works to which they have lawful access in order to carry out a text and data mining analysis will be unenforceable'[73]. However, publishers can adopt 'technical protection measures' that 'reasonably restrict' access to their material to ensure the 'stability and security' of their network.

The exception introduced by the UK reflects a tradition of strong copyright protection. By limiting TDM to non-commercial research the UK exception is of more uncertain application than the proposed Directive. Under the proposed EU Directive, it would be sufficient to demonstrate that the research organisation benefiting from the copyright exception operates on a non-profit basis or in pursuit of a recognised public interest mission. By contrast, under UK law, researchers would have to demonstrate the non-commercial nature of the research itself. In the absence of case law on the topic, the application of the exception remains unclear in cases where a research project involves commercial partners, or could result in a commercial application.

## 3. The 'Loi pour une République Numérique'

On 28 September 2016 France passed a law granting an exception from copyright for TDM. The adoption of the so-called Digital Republic law (Loi pour une République Numérique)[74], also known as Loi Lemaire, follows a long debate and considerable controversy. Its development includes a

[69] IPO. Impact Assessment. (2012)

[70] Intellectual Property Office (UK) Viscount Younger of Leckie. Press release. New exceptions to copyright reflect digital age (2014)

[71] IPO. Copyright, Designs and Patents Acts 1988, s29a

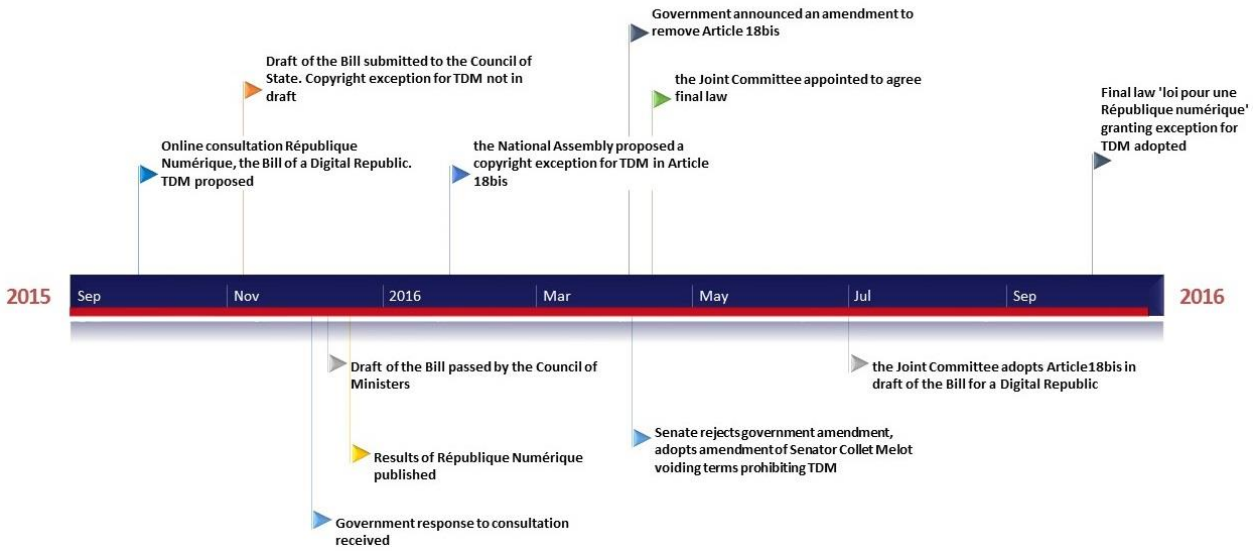
[72] IPO. Copyright, Designs and Patents Acts 1988. s29

[73] IPO. Exceptions to copyright - Research. (2014)

[74] Projet de loi pour une République numérique

nationwide online consultation held in September 2015 which saw the participation of 21,330 contributors who voted nearly 150,000 times and submitted more than 8,500 arguments, amendments and proposals for new articles at the website www.republiquenumeriques.fr. The key events leading to the adoption of an exception are summarised in the timeline below:

**Figure 2 Timeline - The adoption of a TDM exception in France**




The key provisions of the French exception are summarised in former article "18bis" of the Loi Lemaire (now article 38). Article 38 also appends a new alinea 10 in Law 122-5 as follows:

> **'Les copies ou reproductions numériques réalisées à partir d'une source licite, en vue de l'exploration de textes et de données incluses ou associées aux écrits scientifiques pour les besoins de la recherche publique, à l'exclusion de toute finalité commerciale. Un décret fixe les conditions dans lesquelles l'exploration des textes et des données est mise en oeuvre, ainsi que les modalités de conservation et de communication des fichiers produits au terme des activités de recherche pour lesquelles elles ont été produites; ces fichiers constituent des données de la recherche'**[75]

Moreover, the second paragraph of the same article (amending a new alinea 5 in Law 142-3) further states that:

> **'Les copies ou reproductions numériques de la base réalisées par une personne qui y a licitement accès, en vue de fouilles de textes et de données incluses ou associées aux écrits scientifiques dans un cadre de recherche, à l'exclusion de toute finalité commerciale. La conservation et la communication des copies techniques issues des traitements, au terme des activités de recherche pour lesquelles elles ont été produites, sont assurées par des organismes désignés par décret. Les autres copies ou reproductions sont détruites' (ibid).**

[75] Projet de loi pour une République numérique

The scope of the French copyright exception thus presents a series of limitations, some of which are in line with the UK exception while others go further. Below is a summary of the key limits to the exception:

(1) TDM can be performed only from a corpus whose contents were acquired lawfully
(2) …and which contains either any kind of text, or data included in or associated with scientific writings
(3) TDM can be performed only for the needs of public research
(4) Any commercial purpose is excluded
(5) Conservation and treatment of technical copies for the research activities are carried out by bodies designated by decree. Other copies or reproductions are to be destroyed[76]

In each case, there are likely to be disputes over the interpretation and scope of the exception, and differing views on its applicability in specific cases. In France, much will depend on the organisations given permission to conserve and communicate copyrighted material. We understand that this is likely to include the Bibliothèque nationale de France (BnF)[77] and organisations involved in ISTEX[78] (see section 6). The mechanisms by which these organisations fulfil this role will have a profound impact on the national TDM infrastructure, and will determine how easily French researchers are able to access and re-use content for mining purposes.

[76] Projet de loi pour une République numérique. TEXTE ÉLABORÉ PAR LA COMMISSION MIXTE PARITAIRE

[77] Please see http://www.bnf.fr/fr/acc/x.accueil.html for more information

[78] Please see http://www.inist.fr/?Istex&lang=fr for more information

# Appendix C – Abbreviations and glossary

| | |
|---|---|
| **ANF** | Agence nationale de la recherche. The French National Research Agency tasked with funding scientific research. |
| **API** | Application programming interface. A set of functions and procedures to create an application to extract features or data of an operating system, web site, application, or other service. |
| **BnF** | Bibliothèque nationale de France. The National Library of France, a repository of all materials published in France. Located in Paris, France. |
| **CCC** | Copyright Clearance Centre, Inc. is a U.S. company that provides collective copyright licensing services for corporate and academic users of copyrighted materials. Founded in 1978 and headquartered in Danvers, Massachusetts. |
| **CNIL** | Commission nationale de l'informatique et des libertés. The French National Commission on Informatics and Liberty. An independent French administrative regulatory body that ensures that data privacy law is applied to the collection, storage, and use of personal data. |
| **CNRS** | Centre national de la recherche scientifique. The French National Centre for Scientific Research, the largest governmental research organisation in France. |
| **Couperin** | The Couperin Consortium negotiates the purchase of digital resources on behalf of higher education and research institutions in France |
| **Crossref** | Collaborative reference linking service to make content easy to find, link, cite and assess. It holds no full text content, but makes linkages through Crossref Digital Object Identifiers tagged to article metadata supplied by the participating publishers. The end result is an efficient, scalable linking system through which a researcher can click on a reference citation in a journal and access the cited article. It is a not-for-profit membership association operated by the Publishers International Linking Association, Inc. with offices in the UK and US. |
| **EBI** | The European Bioinformatics Institute. The institute makes the world's public biological data freely available to the scientific community via a range of services and tools, perform basic research and provide professional training in bioinformatics. Part of the European Molecular Biology Laboratory (EBI-EMBL). |
| **GATE** | General Architecture for Text Engineering. A computer architecture for a broad range of Natural Language Processing tasks, available under the GNU Public License. |
| **Github** | Web-based Git repository hosting service |

| | |
|---|---|
| **INA** | Institut national de l'audiovisuel. A repository of all French radio and television audiovisual archives. Also it provides free access to archives of countries such as Afghanistan and Cambodia. Headquarters in Bry-sur-Marne, Paris, France. |
| **INRA** | Institut national de la recherche agronomique. The French national institute for agricultural research. Founded in 1946, it is a French public research institute under the joint authority of the Ministries of Research and Agriculture. |
| **ISTEX** | Initiative d'excellence en Information scientifique et technique. The French Institute for Technical and Scientific Information is a vast program of acquisitions of scientific resources aimed at creating an international-standard digital library which can be accessed by all members of higher education and research institutions. |
| **Jisc** | The UK higher education, further education and skills sectors' not-for-profit organisation for digital services and solutions. |
| **Jisc Collections** | The negotiation and licensing service that supports the procurement of digital content for higher education and research institutions in the UK. |
| **NaCTeM** | The National Centre for Text Mining. Operated by the University of Manchester, UK, it provides text mining services in response to the requirements of the UK academic community. |
| **RightFind for XML** | Content workflow solution by Copyright Clearance Center, Inc. Allows users to get full-text articles from company subscriptions, Open Access and internal documents or purchase articles via document delivery. Also provides rights information to the user. |