



**Standardisation in the area of innovation and
technological development, notably in the field of**

Text and Data Mining

Report from the Expert Group

EUROPEAN COMMISSION

Directorate-General for Research and Innovation
Directorate B — Innovation Union and European Research Area
Unit B.1 — Innovation Union Policy

Contact: Peter Dröll

E-mail: RTD-Innovation-Policy-B1@ec.europa.eu
RTD-PUBLICATIONS@ec.europa.eu

European Commission
B-1049 Brussels

**Standardisation in the area of innovation and
technological development, notably in the field of**

Text and Data Mining

Report from the Expert Group

The Expert Group was Chaired by Professor **Ian Hargreaves** (Cardiff University, United Kingdom) with members Dr **Lucie Guibault** (University of Amsterdam, the Netherlands), Dr **Christian Handke** (University of Amsterdam and Erasmus University, the Netherlands), Professor **Peggy Valcke** (KU Leuven, Belgium) and economist **Bertin Martens**¹ (JRC, IPTS, Seville). They were supported by Dr **Ros Lynch** (Department for Business Innovation & Skills, United Kingdom) as rapporteur. The Expert Group also thanks Dr **Sergey Filippov**, Assistant Professor of Innovation Management at Delft University of Technology and Non-resident Fellow of the Lisbon Council, for research assistance.

Directorate-General for Research and Innovation

2014

¹ As the Commission's in-house science service, the Joint Research Centre's mission is to provide EU policies with independent, evidence-based scientific and technical support throughout the whole policy cycle.
<http://ec.europa.eu/dgs/jrc/>

***EUROPE DIRECT is a service to help you find answers
to your questions about the European Union***

Freephone number (*):
00 800 6 7 8 9 10 11

(*) Certain mobile telephone operators do not allow access to 00 800 numbers
or these calls may be billed

LEGAL NOTICE

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use which might be made of the following information.

The views expressed in this publication are the sole responsibility of the author and do not necessarily reflect the views of the European Commission.

More information on the European Union is available on the Internet
(<http://europa.eu>).

Cataloguing data can be found at the end of this publication.

Luxembourg: Publications Office of the European Union, 2014

ISBN 978-92-79-36743-4
doi 10.2777/71122

© European Union, 2014
Reproduction is authorised provided the source is acknowledged.

Image © Login, image 52925488, 2014. Source: Fotolia.com

Executive summary

Text and data mining (TDM) is an important technique for analysing and extracting new insights and knowledge from the exponentially increasing store of digital data ('Big Data'). It is important to understand the extent to which the EU's current legal framework encourages or obstructs this new form of research and to assess the scale of the economic issues at stake.

TDM is useful to researchers of all kinds, from historians to medical experts, and its methods are relevant to organisations throughout the public and private sectors. Because TDM research technology is not prohibitively expensive, it is readily available to lone entrepreneurs, individual post-graduate students, start-ups and small firms. It is also amenable to playful and highly speculative uses, enabling research connections between previously unconnected fields. There is growing recognition that we are at the threshold of the mass automation of service industries (automation of thinking) comparable with the robotic automation of manufacturing production lines (automation of muscle) in an earlier era. TDM will be widely used to provide insights in the re-design of this digital services economy.

When it comes to the deployment of TDM, there are worrying signs that European researchers may be falling behind, especially with regard to researchers in the United States. Researchers in Europe believe that this results, at least in part, from the nature of Europe's laws with regard to copyright, database protection and, perhaps increasingly, data privacy. In the United States, the 'fair use' defence against copyright infringement appears to offer greater re-assurance to researchers than the comparable copyright framework in Europe, which relies upon a closed set of statutory exceptions. Recent court decisions, for example in the ten-year old 'Google Books' case, appear to confirm this. The US has no equivalent of Europe's database protection laws.

In Europe, there are signs of a response among publishers to encourage wider use of TDM. Scientific publishers have recently proposed licensing terms designed to make TDM of their own archives easier, but many researchers dismiss these efforts as insufficient, arguing that 'the right to read is the right to mine' and that effective research demands freedom to mine all public domain databases without restriction. These pressures from researchers have increased as a result of a growing move to 'Open Access' scientific publishing in Europe and elsewhere. The UK and Ireland have already committed themselves to more permissive copyright rules with regard to TDM.

Stakeholders

An overview of the debate about TDM among stakeholders draws attention to the polarisation of views between publishers (especially of scientific journals) and scientific researchers, but notes that relevant communities of interest extend way beyond these groups to include heritage institutions, technology firms, data management companies, pharmaceuticals, newspapers, healthcare providers, advertising agencies and many more. Any organisation seeking to provide a bespoke service to its customers will potentially have an interest in TDM.

It is difficult to estimate accurately the level of TDM activity taking place in Europe, though it would appear to be limited in some fields. A small study conducted by the Lisbon Council among European academics mainly in the social sciences found that

few were aware of or used TDM themselves.² In other fields, such as computational linguistics, TDM is said to account for almost 30% of all research projects. Some publishers report little interest in TDM; others report signs of growth. Researchers suggest this may reflect problems of data access, time-consuming procedures, legal uncertainties and shortages of sufficiently skilled researchers.

Traditional publishers distinguish between 'access' and 'mining', arguing that they are two different activities that require their own licence and may bring with them different terms and conditions. Providing researchers with ongoing, reliable access to high quality content for text and data mining is said to involve a significant investment in validation, correction and refinements to content, plus investment in systems to hold that content in a secure manner. At the same time, there is some acceptance among scientific publishers that the present arrangements are inefficient and costly and would not scale if demand for TDM were to grow as predicted.

Following on from the EU's 'Licences for Europe' process traditional publishers have argued for a 'market solution' based upon collaboration between the various parties. Reed Elsevier recently announced that researchers at academic institutions can use their online interface (API) to batch-download documents in computer-readable XML format, with a limit of 10,000 articles per month. PLOS, on the other hand, recently announced that it would require authors to sign a data availability statement that would guarantee, unless in few exceptional cases, that all the data used in a publication is publicly accessible to anyone at the moment the article is published.

Many researchers, however, do not believe that licensing can solve the problems they face. They call for a revision of copyright law, perhaps in the form of an exception for TDM along the lines proposed in the UK and Ireland, along with reform of EU database law.

Researchers and publishers also disagree about a number of the technical difficulties involved in improving the conditions for TDM and related costs. The growth of Open Access publishing has tended to support the argument that researchers using TDM should not face restrictions. This argument has been supported in the context of the EU's Horizon 2020 strategic research and innovation framework. It is acknowledged that the changes in the technologies which support research present serious questions for the business models of some publishers.

Economic issues

In thinking about copyright, economic policy-makers aim for a welfare-maximising balance between benefits for users and incentives for rights holders. There is a severe lack of empirical evidence upon which to base such calculations, though the theoretical issues are relatively well understood. These rest upon striking the right balance between incentivising the production of 'works', whilst avoiding 'deadweight' welfare losses, for example through excessive transaction costs.

Solid evidence about the prevalence of TDM is scarce, but what evidence there is suggests strong rates of growth from a low base in the last five years. Based upon an analysis of citations which mention data mining in the title of a publication, US researchers appear to be more active than in other countries, though there are also disparities between European countries.

² Cited in Filippov, *Mapping the Use of Text and Data mining in Academic and Research Communities in Europe*. Lisbon Council, Brussels, forthcoming.

Based upon assumptions in a range of studies, estimates are made of the potential value of TDM to Europe's economy, assuming an increase in researcher productivity of 2 per cent and consequent growth in the volume of research and its associated benefits. On conservative assumptions (a narrow definition of the scope for TDM), a GDP gain in Europe 'of the order of magnitude of tens of billions of Euros' appears feasible.

A discussion of market failure and the shortfall in competitive TDM in Europe considers three reasons why the transformative and economically valuable secondary use of copyright works (as exemplified by TDM) may be suboptimal. These factors are: transaction costs, strategic behaviour by copyright holders and externalities. In considering the potential economic consequences of changes in the law governing TDM, five definitions of the boundaries of TDM are considered in order to address the critical economic question of the extent to which any given legal reform will or will not adversely affect the supply of new works, in ways likely to affect the balance of welfare.

In considering various possible forms of legal exception from copyright and database law for text and data miners, the argument is made that from an economic perspective it makes little sense to propose a distinction between commercial and non-commercial TDM. A well-designed copyright regime should provide appropriate stimulus for all types of research and, at the same time, an appropriate level of protection for all rights owners. Once this balance has been reached, there is no reason to distinguish between commercial and non-commercial research.

Legal issues

This section asks whether legal barriers impede the conduct of TDM for research purposes and, if so, how these barriers might be alleviated in the light of the current European legal framework, taking the interests of all stakeholders into account. A range of potential reforms is discussed.

A description is offered of the application of intellectual property laws relevant to TDM in the United States and four other countries. In the US, it is judged reasonable to assume that copying acts by American TDM researchers for the purpose of extracting non-expressive metadata could be considered fair use under US law. Under Canadian law, TDM activities would likewise probably qualify as fair dealing. Australia's legal regime appears to be more restrictive than in North America. The picture is less clear cut in Japan and Israel, though in both these countries there have been legal changes which may be helpful to researchers using TDM.

The extent to which TDM in Europe is facilitated by any existing exceptions to either EU copyright or database law appears unclear. The application of a copyright and database exception relating to teaching or scientific research is optional and has not been implemented at all in some Member States. This has contributed to uncertainty in the European scientific research community.

Encouraging TDM for research purposes without fear of infringing IP rights could be achieved in a number of ways: through an adjustment of licensing practices; through a revised, normative interpretation of the 'reproduction right'; through the introduction of a new exception in copyright and database laws, or through the adoption of an 'open norm' designed to guide the courts to take a more flexible view of what users are permitted to do. Should an exception be introduced in the European legal framework, the legislator would also need to consider whether to ensure that it cannot be over-ridden through the enforcement of restrictive contractual clauses or technological protection measures.

An approach based upon licensing alone would probably be insufficient to allow TDM to take place in all instances where it would be socially desirable because of uneven levels of access, high transaction costs and patchy availability of works covered by a creative commons licence.

A more promising route could involve reconsideration of the right of reproduction in copyright law, along with the right of extraction in the database regime. These have traditionally been subject to increasingly broad interpretation, but the need to boost TDM in Europe provides impetus to consider a change of emphasis. This would involve the legislator adopting a 'normative' approach, designed to ensure that protection is supported by the courts only for acts of reproduction or extraction that entail 'expressive' exploitation of the rights-protected material. This would put TDM's non-expressive and socially beneficial mechanical sifting of data beyond successful challenge in the courts. Such a shift could be achieved through an interpretation instrument issued by the European legislator, accompanied by a re-assessment of the Database Directive, building upon the European Commission's own highly critical evaluation report in 2005.

A third alternative would be to introduce a new exception in copyright and the database law. This might take one of two forms: an exception specifically permitting TDM for the purpose of research or an open norm. The first would provide more immediate clarity; the second would offer more flexibility in a fast changing technological environment. An 'open norm' approach could involve a re-balanced interpretation of the Berne Convention's Three Step Test.

Finally, two areas of legal discussion beyond IP law are considered. The first concerns demands to resist the 'monopolisation of information' by major holders of data, potentially through the operation of competition law. Among the ideas discussed is the call for a more general regime of mandatory openness and interoperability (with open standards) in online environments, designed to prevent a major data holder (one might think of Facebook, Twitter, Google or other online players) 'from erecting a fence around its piece of the information commons.'

The second area of non-IP law concerns data privacy, where already strong European laws protecting individual privacy stand to be strengthened by the draft Data Protection Regulation currently under consideration. This draft legislation includes a provision explicitly permitting the processing of even sensitive personal data for the purposes of historical, statistical or scientific research, subject to certain safeguards. It has been argued, however, that the draft legislation will prove problematic for TDM, because mining requires sweeping assemblies of data and an exploratory, iterative approach to research goals. Some researchers argue for a shift of regulatory attention away from data *collection* and towards the way that data and knowledge based on data are *used or abused*.

Conclusions

From the analysis in this paper, we can draw the following analytical conclusions about TDM and the challenge it presents to policymakers in Europe:

- Text and data mining is an important research technique which is certain to become more important as researchers acquire the skills and the technology to address and investigate datasets of increasing size, complexity and diversity in all media: text, numbers, images, audio files and in any other form.
- TDM represents a significant economic opportunity for Europe. Prolific use of TDM would add tens of billions of Euros in value to the EU's aggregate GDP.

This would result chiefly from higher productivity among researchers and from the effects ('externalities') of increased levels of research.

- At present, the use of TDM tools by researchers in Europe appears to be lower, and probably significantly lower, than is the case in the United States and some other countries in the Americas and Asia. This probably reflects, among other factors, disadvantages created by the European legal framework with regard to TDM.
- The European legislator needs to re-consider and reform the EU's legal framework with regard to copyright, database protection and possibly data privacy, in order to support the international competitiveness of Europe's research base.
- There is a serious risk that Europe's relative competitive position as a research location for the exploitation of 'Big Data' will deteriorate further, if steps are not taken to address the issues discussed in this report. The results of this might well include a loss of talent and a loss of investment to more favourable research locations.

In response to this analysis, the Expert Review group proposes three action points:

1. We welcome initiatives to make licensing of works for the purpose of text and data mining easier. In the short term, these will add value to the economy and help to build the skills-base and culture necessary for successful 'big data' research in the digital economy. This activity, however, should be seen as a prologue to legal reform, not an end in itself.
2. A specific and mandatory exception to remove text and data mining for scientific purposes from the reach of European copyright and database law should be drafted. This should be regarded as a short-term amelioration, in the event that our third proposal, below, cannot make timely progress.
3. The best approach to reform, aimed at securing a competitive legal framework for European research, is to establish a durable distinction in European law between copyright's longstanding and legitimate role in protecting the rights of authors of 'expressive' works and copyright's questionable role in the digital age of presenting a barrier to modern research techniques and so to the pursuit of new knowledge. This initiative should be at the heart of a new copyright directive in Europe, following the consultations currently being undertaken by the European Commission. The legal analysis in this report offers more than one route via which a reform of this kind might be pursued; for example by introducing a suitable 'interpretative instrument' into a new Copyright Directive. We also urge the legislator, including the European Parliament, to ensure that the currently proposed reform of Europe's data protection laws avoids the unintended consequence of creating further impediments to the work of scientific researchers. We make these recommendations in the interests of the international competitiveness of the European Union's research base.

Table of Contents

1. INTRODUCTION	8
1.1 Definitions	9
1.2 Big Data	10
1.3 International comparisons.....	11
1.4 Licensing versus legal reform	12
2. STAKEHOLDER VIEWS.....	14
2.1 Facilitating TDM access	14
2.2 Legal rights to undertake TDM	19
2.3 Technological challenges.....	20
2.4 Cultural challenges.....	21
3. ECONOMIC ISSUES	24
3.1 Basic economic considerations	24
3.2 Empirical evidence	26
3.3 Economic consequences of legal reform	31
3.4 Market failure: what prevents competitive TDM in Europe?	34
3.5 The scope for special copyright arrangements for TDM	38
3.6 An exception for TDM for non-commercial research only	40
4. LEGAL ISSUES	42
4.1 TDM outside Europe	43
4.2 TDM and European Intellectual property protection	47
4.3 TDM and the current research exception	49
4.4 Making room for TDM activities under IP law.....	50
4.5 Licensing solutions	51
4.6 Statutory exception.....	53
4.7 Open Norm	56
4.8 Accessing non-protected databases	57
4.9 Privacy issues	60
5. CONCLUSIONS	64
5.1 Licensing	64
5.2 An exception favouring text and data mining	65
5.3 A strategic reform of copyright and data-base law	66
BIBLIOGRAPHY	68
APPENDIX: An exploration of Google Scholar data	69

1. Introduction

There is widespread agreement that the effective harnessing of digital communications technologies is important to the performance of advanced economies, such as those of the European Union (EU). Text and data mining, the subject of this report, offers a significant set of techniques for exploiting the research potential of these technologies.

Advanced economies are increasingly dependent upon investment in intangible rather than fixed assets³ and they rely heavily for innovation upon smaller firms which successfully deploy these technologies. The intangible assets in which companies in advanced economies invest, such as brand, product design, training and software development are, to a considerable extent, the subject of protection by the laws governing intellectual property. A comprehensive and recent study suggests that IP-intensive industries accounted for 35 per cent of all the jobs created in the EU between 2008 and 2010, along with 39 per cent of total economic output and 90 per cent of exports.⁴ At the same time, many of these IP-intensive industries are experiencing business model disruption from digital technologies, highlighting painful trade-offs between established and new players in many markets.

Navigating these tensions in order to preserve the legitimate role of copyright and other IP rules, whilst also promoting successful innovation and enhanced productivity, has proved elusive in Europe in the last decade. The EU's productivity shortfall is well documented and recognised in the goals of the EU's Horizon 2020 research and innovation framework,⁵ which states its overarching priority as "exiting the economic crisis through sustainable growth." The programme's 'future and emerging technologies' theme points to the need 'to promote and support the emergence of radically new technology areas that will renew the basis for future European competitiveness and growth.'

These are background points in the pivotal debate concerning the actions needed to stimulate the EU's digital economy by overcoming blockages in markets caused by geographic and legal fragmentation in order to establish a 'digital single market,' which builds upon the single market that underpinned EU prosperity in the late 20th century.

These same points also provide crucial context for the subject of this expert review: the development of text and data mining (TDM) within the European Union. TDM is a tool potentially capable of stimulating innovation in many business sectors and

³ See, for example, Nesta's Innovation Index: <http://www.nesta.org.uk/publications/innovation-index-2012>; and OECD 2013: *Intangible assets, resource allocation and growth: a framework for analysis*: <http://dx.doi.org/10.1787/5k92s63w14wb-en>; and Hargreaves: *Digital Opportunity: a review of intellectual property and growth*, UK IPO 2011.

⁴ *Intellectual Property Rights Intensive Industries: contribution to economic performance and employment in the European Union*: European Patent Office and the Office for Harmonisation in the Internal Market. September 2013.

⁵ ec.europa.eu/programmes/horizon2020/

across the public sector, whilst at the same time raising the productivity of Europe's researchers and contributing to the growth of Europe's GDP.

1.1 Definitions

Text and data mining involves the deployment of a set of continuously evolving research techniques which have become available as a result of widely distributed access to massive, networked computing power and exponentially increasing digital data sets, enabling almost anyone who has the right level of skills and access to assemble vast quantities of data, whether as text, numbers, images or in any other form, and to explore that data in search of new insights and knowledge.⁶

TDM is important to researchers of all kinds. A historian with the necessary skills and an accessible digital archive can check the frequency with which a particular set of terms was used in the first half of the 19th century, compared with the second half. Analysis of vast quantities of video is crucial to research in meteorology and police forensics. A researcher in political economy can analyse the incidence and meaning of the word 'digital' in the work of the EU. Retailers can combine their knowledge of shoppers' spending patterns with analysis of their leisure time and health. A medical researcher into Alzheimer's disease may cross-examine unprecedented quantities of neurological and lifestyle data from patient records and investigations in many territories. Genetic studies and astronomy are among the areas of science which have already benefited significantly from these still very new and developing techniques. In short, TDM-based research plays a role in almost every area of human life, from banking, government and newspaper publishing to advanced manufacturing and advertising.

Because TDM research technology is not itself prohibitively expensive, it is readily available to lone entrepreneurs, individual post-graduate students, start-ups and small firms. It is also amenable to playful and highly speculative uses, seeking to apply knowledge in one field (such as human or animal neurology) to others where this would not previously have been thought feasible (such as music, games or the design of furniture and cars). In an emerging world where many objects are connected to each other (via the 'Internet of Things') the rate of increase in the quantity of analysable data will continue to accelerate. This data makes possible new products and services and even entirely new zones of human service provision, such as technologies which enhance personal performance, sometimes called 'transhumanism' or 'Humanity 2.0'. More mundanely, but significantly, there is growing recognition that we are at the threshold of mass automation of our service industries (automation of thinking) comparable with the robotic automation of manufacturing production lines (automation of muscle) in an earlier era. TDM will be widely used to provide insights in the re-design of this digital services economy.

⁶ This definition accords broadly with the one proposed by the Publishing Research Consortium (2013): 'Data mining is an analytical process that looks for trends and patterns in data sets that reveal new insights. These new insights are implicit, previously unknown and potentially useful pieces of information. The data, whether it is made up of words or numbers or both, is stored in relational databases. It may be helpful to think of this process as database mining or as some refer to it 'knowledge discovery in databases. Data mining is well established in fields such as astronomy and genetics.'

1.2 Big Data

All of these activities, along with countless others, involve 'Big Data'. It is said to be true that every day humans create 2.5 quintillion bytes of data and that 90 per cent of this data has been created in the last two years. Social media sites, smartphones and other consumer devices including PCs and laptops have allowed billions of individuals around the world to contribute to this stock of data. Millions of networked sensors are being embedded in devices such as mobile phones, smart energy meters, automobiles, and industrial machines that sense, create, and communicate data. The volume of this incomprehensibly large data store is forecast to double in size every three years.⁷

McKinsey Global Institute estimated in 2012 that the US healthcare industry alone could generate \$300bn in value every year from an efficient and creative use of Big Data. Deployment of services based upon analysis of personal location data was estimated to generate £600bn in consumer surplus. Economists, however, have not been able to reach settled judgments on the scale of the economic impact of this explosion of advanced data analytics, even as they debate its far-reaching impact upon wealth disparities, labour markets, innovation and economic growth.⁸ One reason for this lack of clarity, according to McKinsey Global Institute, is the uncertainty attaching to data access rights, arising from a potential misalignment of stakeholder incentives and so resulting in market failures for the sharing or trading of data.⁹

The definitions of 'data' and 'research' implied in these examples are necessarily and deliberately broad. Research today takes many forms, typically involving multiple disciplines. Some research, as has always been the case, creates new data, but today's researchers also have unprecedented ability to build upon past knowledge.

'Scraping' the World-wide web for data is today a familiar activity for the digitally literate researcher. Data brokerage firms gather this and other information and sell it in bundles in the commercial marketplace. Meanwhile, the results of academic research continue to be shared, to a great extent, through scholarly articles, published in peer-reviewed journals, most of them now available on-line. One estimate suggests that there are today over 50 million such articles in existence.¹⁰ All of this makes it decreasingly possible for any human researcher, or even a substantial research team, to consider all of the potentially relevant literature and data. That is why text and data mining is such a hot topic within the academic research community. All researchers want access to the full potential of the 'big data' mine. Nor can researchers in one country accept that researchers elsewhere have superior access to these tools.

⁷ See: <http://www.ibm.com/software/data/bigdata/what-is-big-data.html>

⁸ See, for a flavour of this debate: Brynjolfsson and McAfee: *The Second Machine Age: work, progress and prosperity in a time of brilliant technologies*. Norton, 2014; Wolf: *If robots divide us, they will conquer*. Financial Times February 4, 2014; The Economist: *Coming to an office near you: what will today's technology do to tomorrow's jobs?* January 18, 2014.

⁹ McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity, at p.108;

http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

¹⁰ *Article 50 million: an estimate of the number of scholarly articles in existence*. Learned Publishing, 23 (3): 258-263. Cited in Filippov, *Mapping the Use of Text and Data mining in Academic and Research Communities in Europe*. Lisbon Council, Brussels, forthcoming.

1.3 International comparisons

This raises questions which lie at the heart of this expert review: how well is Europe doing in encouraging text and data mining? If it is falling behind, what can be done to improve this state of affairs?

In our terms of reference we were asked to consider whether standard-setting is an issue which merits attention with regard to developing a more effective approach to text and data mining, but our exploration of this point with stakeholders, among the relevant literature and in our own experts' examination of the legal and economic issues did not encourage us to spend too much time on this line of inquiry. Although standards are an important device in established markets for technically complex products and services, the TDM marketplace has not yet settled to a point where standard-setting offers a ready opportunity to support increased value.

It is not standards which Europe's researchers want to discuss. Rather, their concern is focused upon the impediments many say they face in exploiting 'big data' using text and data mining, particularly in comparison with their colleagues in the United States, but also in some other countries in the Americas and Asia, including Canada, Singapore, Japan and South Korea. These impediments arise, they say, from aspects of European copyright law; from the EU's so-called 'sui generis' law of 1996 protecting the contents of databases, and, perhaps, from Europe's currently shifting legal framework with regard to data privacy.

Copyright comes into play because text and data mining begins with the unavoidable organisation of data so that it can be analysed. It is the subject of fierce debate whether, for researchers, this act of 'organisation' amounts to copying within the meaning of copyright law. In Europe, some Member States have already adopted an exception or limitation to copyright rules applying generally to academic research, but this exception is both uneven in its application and less permissive than the legal regime in the United States, where the 'fair use' defence appears to offer significantly greater comfort to researchers about what they can and cannot do without fear of provoking successful legal action from rights holders. With its reference point of the First Amendment to the US Constitution, forbidding any abridgement of the right to free expression, and its explicit reference to scholarly research in its 'fair use' doctrine, American jurisprudence in copyright continues to evolve in a more permissive direction, from the point of view of researchers. In the ten-year-old 'Google Books' case, for example, which has set the Silicon Valley technology giant against US authors and publishers, the most recent and high level legal ruling in late 2013 has ruled in favour of Google, agreeing that Google's indexing work qualifies as 'transformative' content. The judgment also refers to freedom of expression and draws specific attention to the importance of text and data mining.¹¹

Further complications, and therefore impediments to TDM, arise from the workings of the EU's 1996 Database Directive, which was designed to boost growth in the European database industry by offering protection for investments in databases of a kind unavailable in the US or elsewhere in the world. A European Commission

¹¹ See, for example: <http://www.wired.com/threatlevel/2013/11/google-books/>

review of this directive in 2005 concluded that the disparity in growth between the EU and US industries since the directive had moved further in favour of American database companies.¹² In spite of this, the directive remains in place.

Then there is a bundle of controversial issues arising from concerns about data privacy and protection, currently leading to new policy initiatives in Europe, which may cause further divergence between the European and American landscape for text and data mining. This follows high level tensions over access to mobile phone calls and other data by American intelligence agencies. One likely impact is that data held in North America, including data of European origin, will attract less rigorous levels of protection compared with data held in Europe.¹³ This may reflect wholly legitimate European sensitivities about data privacy, which go beyond this review's terms of reference. We merely note that this may create a further obstacle to the competitive deployment of text and data mining by Europe-based researchers.

In our detailed examination of these issues, we seek first to describe the 'stakeholder' debate as it stands, drawing upon debates and statements from those who believe they have much to gain or to lose from text and data mining. We then consider the potential economic issues at stake, before turning to the legal issues, where we begin by asking whether the legal and operational status quo is a viable option for Europe, given these economic calculations.

1.4 Licensing versus legal reform

In practice, there are few voices defending the status quo as such; a clear indication of the timeliness of the decision to commission this review. When TDM first emerged in the 1990s, scientific publishers resisted it on the grounds that it was of minority interest and did not appear to be good for their own businesses, offering no clear, new revenue stream, imposing potential additional costs on database management and adding to the risk of online piracy. In recent months, however, traditional publishers have shifted position, following encouragement from the European Commission's 'Licences for Europe' initiative and pressure from academics, who create the material scientific publishers sell.

These pressures for change have also been accentuated by a growing 'Open Access' model of scientific publishing within and beyond Europe. Supporters of Open Access, including open access publishers, argue that since most scientific research is publicly funded, it ought as a matter of principle to be freely available to anyone to read or to mine, using computer algorithms. A 2013 study estimates that more than 40% of scientific peer reviewed articles published worldwide between 2004 and 2011 are now available online in open access form¹⁴.

This background also helps explain why some European countries including the UK and Ireland, have committed themselves to specific reform of the rules governing

¹² European Commission 2005. *First evaluation of Directive 96/9/EC on the legal protection of databases*. DG Internal Market and Services Working Paper.

http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report.

¹³ Though American privacy law also offers greater privacy protection to American citizens and corporate bodies than it does to others.

¹⁴ http://europa.eu/rapid/press-release_IP-13-786_en.htm

TDM.¹⁵ Reed Elsevier, one of the world's largest scientific publishers, recently proposed new licensing terms for access to TDM. This initiative has been welcomed in some quarters, but many researchers argue that only an explicit exemption from copyright for TDM as a technique will foster a TDM culture and practice on the scale needed. Campaigners argue that 'the right to read is the right to mine' and so resist the publishers' claims to additional contractual terms, charges or controls for text and data mining.

Others go further and argue that TDM is of such pivotal importance to research and of such high economic value that it needs to be readily available not only to academic researchers, but also to scientific research conducted in the commercial arena. Economic arguments suggest that the welfare gains from commercial TDM would greatly exceed those available from non-commercial TDM. This argument also holds that making a distinction in law between 'commercial' and 'non-commercial' research would be difficult if not impossible, especially in a time when academics are encouraged, increasingly, to collaborate and 'co-create' with business.

From here, the argument for reform takes a different shape, examining a more general solution than a tightly drawn exception for copyright and data base law in the form of an amendment to the basic definition of the 'reproduction right' in copyright designed to distinguish between copyright's core purpose in motivating artistic works and its acquired effect in the digital age of obstructing use of some digital technologies, such as TDM.

Finally, it should be added here that in focusing our attention upon legal and economic issues, we do not in this report consider in detail other factors which no doubt provide part of the explanation for Europe's TDM deficit: such as skills, cultures of innovation, logistics and digital infrastructures. These are all discussed in a recent OECD study¹⁶, which speaks of 'a shift towards a data-driven socio-economic model' where 'data are a core asset which can create a significant competitive advantage and drive innovation, sustainable growth and development.' It is beyond dispute, however, that a clear and predictable legal framework with regard to TDM is of the utmost importance to European researchers' text and data mining activities in the years ahead.

¹⁵ In the UK, the 2011 review known as the 'Hargreaves Review' and in 2013 the Irish Copyright Review recommended legal changes designed to make TDM more available.

¹⁶ OECD 2013: *Exploring data-driven innovation and new sources of growth: mapping the policy issues raised by Big Data*. OECD Digital Economy Paper No 222.

2. Stakeholder views

The issue of text and data mining (TDM) has been hotly debated among stakeholders in the UK and more recently in Europe. On the face of it, these debates seem to be polarised between publishers (mainly journal) and researchers (largely scientific). However the communities of interest go much wider and include cultural heritage institutions, technology firms, data management companies, pharmaceuticals, newspapers, healthcare providers, advertising agencies and many more. In fact, any organisation seeking to provide a bespoke service to its customers will potentially have an interest in TDM.

The timescale for this project did not allow for a full consultation with these communities. Instead the stakeholder views presented here are drawn from responses to the two main consultations run by the UK IPO¹⁷, various papers and opinion pieces published on the subject, and interviews/discussions with a small number of stakeholders in Europe.

2.1 Facilitating TDM access

As indicated in the introduction to this report, we live increasingly in a data-driven world. As more and more data becomes available researchers from all fields need to find better ways of making sense of it. TDM is one of the tools being employed by researchers and data users more generally to maximise the benefits from the explosion in data.

However, it is extremely difficult to estimate accurately the level of TDM activity taking place in Europe though it would appear to be limited in some fields of study. A small study conducted by the Lisbon Council with European academics mainly in the social sciences found that few were aware of or used TDM themselves.¹⁸ In other fields of study TDM is more widely used. Professor van den Bosh at Radboud University, Nijmegen, estimates that “in the field of computational linguistics (or human) language technology, natural language process), TDM accounts for about 25-30% of all research projects...”¹⁹

According to the Association of Learned and Professional Society Publishers (ALPSP) the larger publishers receive less than 10 requests per year to text and data mine, while smaller publishers have not received any requests. From a traditional publisher’s point of view, this suggests that there is little demand for TDM and therefore no market failure to address.

¹⁷ Responses to the Hargreaves review on Intellectual Property and Growth and to the Government consultation on the introduction of an Exception for TDM for research.

¹⁸ Cited in Filippov, *Mapping the Use of Text and Data mining in Academic and Research Communities in Europe*. Lisbon Council, Brussels, forthcoming.

¹⁹ *ibid.*

Others disagree with this view and point to a number of reasons why TDM activity may be restricted. These include:

- Legal uncertainty leading to the fear of being sued
- Inaccessible information silos and difficulties involved in linking such varied data
- Lack of quality tools/applications and appropriately skilled people to use them
- Contacting and negotiating with multiple publishers is time-consuming and costly. According to Jisc²⁰, a UK charity focused upon digital research issues, a single researcher seeking to mine PubMed Central articles on malaria could lose over 60% of their working year at a transaction cost (in terms of time spent) of £18,630 contacting the 1024 journals necessary to obtain access to the 49% of articles not published via Open Access.²¹
- Inability to obtain standardised content from multiple publishers

For most researchers the key issue is being able to mine content for which they already have legal access. Many within this community believe that academic research should be open and access facilitated through Creative Commons and Open Source Licences²². It is felt that traditional publishers are already adequately compensated (through journal subscriptions) and therefore no further payment for mining content is warranted. Many subscribe to the view that 'the right to read is the right to mine.'²³ Traditional publishers however distinguish between 'access' and 'mining', arguing that they are two different activities that require their own licence and may bring with them different terms and conditions. In addition, providing researchers with ongoing, reliable access to high quality content for text and data mining is said, by traditional publishers, to involve a significant investment in validation, correction and ongoing refinements to content, plus investment in systems to hold that content in a secure manner.

Nevertheless there appears to be some acceptance among the scientific publishing community that the present arrangement is inefficient and costly, and importantly would not scale if demand for TDM were to grow. Following on from the 'Licences for Europe' process²⁴ traditional publishers have been developing specific licences and tools to facilitate TDM:

- Science, technology and medical (STM) publishers have issued a declaration²⁵ setting out their commitment to facilitate TDM for non-commercial, scientific research in the European Union. The declaration has so far been signed by 16

²⁰ JISC was formerly an acronym for the Joint Information Services Committee, but Jisc is now the corporate name of a charity.

²¹ Value and Benefits, p. 27-28.

²² The Lisbon Council, op cit.

²³ UK university libraries, for example, pay publishers around £180 million a year on licences for books and journals (mainly online). In 2013 they paid £28 million to Reed Elsevier and over £14 million for access to Wiley journals. Figures provided by RLUK.

²⁴ See <http://ec.europa.eu/licences-for-europe-dialogue/en/content/about-site>

²⁵ http://www.stm-assoc.org/2013_11_12_News_Release_STM_sector_submissions_to_Licenses_for_Europe_Initiative.pdf

publishers who represent approximately 50% of the world's literature in STM, social science and humanities.

- Crossref – the industry content identification and linking platform has developed 'Prospect,' designed especially to facilitate TDM by non-commercial researchers and their institutions. Researchers will be able to select publishers of interest, accept their licence terms and conditions, and then receive a unique Client API token which facilitates access to the publishers' content.
- The UK Publishers' Licensing Society (PLS) is developing PLS Clear – a web portal to guide mainly unaffiliated researchers through the process of securing permissions and access from publishers. It will be launched in 2014.
- Copyright Clearance Center (CCC) – a US based licensing and rights broker piloted a process to make it easier for commercial researchers to gain quick access to full-text content for mining in a centralised manner with a common interface. CCC estimates that if the 5 publishers in the pilot²⁶ were each to negotiate TDM rights, feeds and data standards with corporate users it would require 25 negotiations, with each negotiation typically taking 2-4 months. The 'hub and spoke' product being rolled out later this year significantly reduces the time involved in the process.

A number of researchers and data analysts contacted for this Expert Review, however, do not believe that licensing is the solution and argue instead that the only truly effective solution would be a revision of copyright law, probably in the form of an exception for TDM along the lines of that proposed in the UK. The League of European Research Universities (LERU)²⁷ in its 'Roadmap for Research Data' published in December 2013 argued that "what is needed at a European level is a Fair Dealing exception certainly for the purposes of research, in the EU Copyright and Database Directives to facilitate the sharing and re-use of research data". This will facilitate greater collaboration among European researchers as required by programmes like Horizon 2020. The Open Access Scientific Publishers Association (OASPA) states that one criterion for membership is that a publisher must use a liberal licence that encourages the reuse and distribution of content. The organisation strongly encourages but does not currently require the use of the CC-BY licence wherever possible.²⁸ Professor van den Bosch argues that "Academic research should be open. Licence forms such as Creative Commons for texts and Open Source licences for software are vital to ensure this openness and should be used wherever possible..."²⁹ Paul Keller, Vice Chair of Kennisland, a Dutch think tank, agrees but goes further arguing that "it should be explicitly stated in law that

²⁶ Royal Society for Chemistry, CABI, Wiley-Blackwell, Sage and Nature

²⁷ The 22 members of LERU include: Universities of Amsterdam, Barcelona, Cambridge, Edinburgh, Freiburg, Genève, Heidelberg, Helsinki, Leiden, Leuven, Lund, Milan, Oxford, Pierre & Marie Curie, Strasbourg, Utrecht, Zurich, Paris-Sud, and Imperial College London, University College London, Ludwig-Maximilians-Universität München

²⁸ <http://oaspa.org/why-cc-by/>

²⁹ Lisbon Council, forthcoming, op cit.

technical protection measures and contracts should not override such an Exception.³⁰ We return to these issues in Chapter 4 of this report.

Traditional publishers disagree. They argue that an exception will not afford access and that what is needed is a market solution based on collaboration between the various parties. Wiley believes that “licences are an effective means of providing certainty and clarity both to rights-holders and end-users ... legislation is likely to discourage innovation in the market, will offer little if any certainty to users wishing to access content for TDM purposes, and will not solve any of the significant technology and security issues that need to be addressed in this context.”³¹

Newspaper publishers are also against the introduction of an exception for TDM. The European Newspaper Publishers Association (ENPA) believes “it could be misused by news aggregators and media monitoring companies in order to avoid the necessity of licensing their activities”. Newspaper publishers maintain that licensing for TDM must be done on a case by case basis even for non-commercial research to prevent massive abuse or loss of their archives and the destruction of their business model.³²

Separately, individual publishers are developing their own responses. On 26 January 2014 Reed Elsevier announced that researchers at academic institutions can use their online interface (API) to batch-download documents in computer-readable XML format. Elsevier has chosen to provisionally limit researchers to 10,000 articles per week. These can be freely mined — so long as the researchers, or their institutions, sign a legal agreement including certain conditions.³³ This, along with the licensing conditions being imposed by the publisher has raised concerns among librarians, particularly in France.³⁴ It is however anticipated that other publishers will follow suit.

The research community is divided over the potential benefits of initiatives such as that launched by Elsevier. Richard Walker, spokesman for the Human Brain Project, argued that “it resolves genuine technical issues”.³⁵ Others are less positive. Peter Murray-Ross has urged researchers and their institutions not to sign up for Elsevier’s TDM service arguing that APIs make it hard to mine and that the burden of mining would increase significantly if every publisher came up with a similar process.³⁶ Richard Van Noorden writes that “some scientists object that even as publishers roll out improved technical infrastructure and allow greater access, they are exerting tight legal controls over the way that text-mining is done.”³⁷ Representatives from the Europe Bioinformatics Institute (EBI) believe that the Elsevier approach will not

³⁰ Lisbon Council, forthcoming, op cit.

³¹ Duncan Campbell, Associate Director, Journal Digital Licensing, Wiley, February 2014

³² ENPA written response to the DG Research Expert Group on Standardisation, February 2014

³³ Conditions include: researchers may publish the products of their text-mining work only under a licence that restricts use to non-commercial purposes, can include only snippets (of up to 200 characters) of the original text, and must include links to original content. Researchers must also register their project on Elsevier’s website (<http://www.developers.elsevier.com/cms/index>)

³⁴ <http://f.hypotheses.org/wp-content/blogs.dir/1658/files/2014/02/Communique%CC%81-Couperin-Ne%CC%81gociation-Elsevier.pdf>

³⁵ Richard Van Noorden, *Elsevier opens its papers to text-mining*, Nature, Vol. 506, 6 February 2014

³⁶ Peter Murray-Ross blog – Content Mining: why you and I should not sign up for Elsevier’s TDM service, 3 February 2014.

³⁷ Van Noorden, op cit.

fundamentally change the model but is in effect another way of controlling access for researchers.³⁸

It is too early to say what impact these initiatives will have. However, the National Centre for Text Mining (NACTEM) believes that while there may be some merit in the licensing proposal, it is highly unlikely it will be effective in facilitating text mining. They point to the experience of JISC Collections which had sought to introduce a clause in its model licence to permit TDM activities. Of 17 journal licence agreements negotiated with major journal publishers, 6 publishers accepted the clause, 6 rejected the clause in its entirety and 5 made significant amendments to the clause.³⁹ Erik Ketzan in his presentation to the 4th meeting of the Licences for Europe Working Group on TMD argued that while licensing could be an option in the short term, in the long term legislative measures would be necessary.

Dr Cameron Neylon (PLOS⁴⁰) believes that the outcome could potentially be a complex system where researchers will have to operate through multiple proxies and 'click throughs' to get the information they need. As more and more data is made available and becomes more distributed, a centralised clearing house will not solve this problem though it could be helpful in the short term. In his view an exception in law will enable critical mass to be reached by encouraging more researchers to become involved in TDM and by reducing significantly the friction in the licensing system. However he accepts there will be a lag, and potentially a long one, before researchers fully understand what they can do and ambition grows.

Neylon argues that EU 'sui generis' database rights already cause a stifling effect compared to the status of data and data collections under US law. In his view, the UK and EU run the risk of falling behind in this space and giving significant legal advantages to those operating under US law. A fuller discussion of the US fair use and EU database rights is provided in Chapter 4.

The Irish Copyright Commission believed that there were significant benefits to be gained from a copyright exception in favour of content mining for non-commercial research. The Government therefore proposed that an exception be cast in fair dealing terms.⁴¹

Whether TDM is facilitated by innovative licensing or by an exception to copyright, there may still be a broader access issue to address. At present scientific articles and the underlying data are stored in different repositories in different countries. The European Bioinformatics Institute (EBI) therefore believes that the Commission should also consider what investment is needed to develop the infrastructure to make the data available in a way that will make it easier for researchers to access and mine. As far as the EBI is concerned it would not make sense to create this infrastructure on an individual country basis.

While the focus of much recent policy debate has been on TDM for non-commercial research, there was a strong view expressed by the majority of people (outside publishing) contacted for this project that it would be unwise to consider an

³⁸ The European Bioinformatics Institute is Europe's flagship laboratory for the life sciences. EBI provides freely available data from life sciences experiments covering the full spectrum of molecular biology.

³⁹ JISC Collections response to UK Government consultation, March 2012

⁴⁰ PLOS is a non-profit open access scientific publishing project. See <http://www.plos.org/about/plos/>

⁴¹ Modernising Copyright: the report of the Copyright Review Committee, October 2013

exception for non-commercial research only. Arguments put forward include the fact that the distinction between commercial and non-commercial research is not clear cut; researchers in both academia and industry are reliant on the same data and often share data across institutions and the new market which it is anticipated increased TDM activity would bring could be stifled.

2.2 Legal rights to undertake TDM

A full discussion of the current legal context and the relationship between IP, database rights and the legality of engaging in TDM activities across a number of EU Member States is provided later in this report. In this section we merely report some of the views expressed by stakeholders on the legality (or not) of engaging in TDM for research.

As previously indicated many researchers believe that the current low level of TDM activity derives in part from legal uncertainties. As licence terms are not always clear, many researchers prefer not to engage in TDM lest they should be sued. Dr Huijnen argues that “copyright law severely hampers our research. The fact that we cannot process newspapers (and other types of historical information) of more recent date (less than 70 years old) because of copyright issues is the main reason we, in our research project, cannot speak of ‘big data research ...’”⁴²

In its response to the UK Government consultation on an exception for TDM for non-commercial purposes, Jisc quotes Korn et al⁴³ who argued that TDM discussions “provoke complex IPR and licensing issues specifically compounded by:

- The inherent copyright and/or database rights which might exist in original texts
- The levels of adaptation and processing required to create the derived data
- The intended use of the outcomes

One of the main disagreements appears to centre on the amount of copying being done. To undertake TDM a researcher must access, or arguably make a copy of the articles/data in order to apply the necessary algorithms. The National History Museum argues that this “in no way conflicts with the legitimate interests of the rights owner. Further it argues that it is the facts dispersed throughout the content and relationship between the facts which are of interest to scientific researchers, neither of which are in themselves protected by copyright.”⁴⁴

The UK Parliament’s Business, Innovation and Skills Committee⁴⁵ did not fully accept this argument, believing that “the assertion that copyright does not restrict the use of facts overlooks the point that scientific papers are not merely presentations of fact; they are interpretations of fact which have typically been peer reviewed and

⁴² Lisbon Council research, op cit.

⁴³ <http://www.jisc.ac.uk/media/documents/projects/iprinderiveddatareport.pdf>

⁴⁴ National History response to the UK Government consultation

⁴⁵ The Business, Innovation and Skills Committee conducted an inquiry into the recommendations set out in the Hargreaves Review of Intellectual Property and the Government’s plans for the implementation of its recommendations. See <http://www.publications.parliament.uk/pa/cm201213/cmselect/cmbis/367/367.pdf>

edited, with a substantial contribution to the editing process usually deriving from publishers.” It held that publishers have a legitimate argument that a licence for human readership differs from one that permits wholesale computerised reading in legal and technical terms.

In contrast, the Australian Industry Information Association (AIIA) suggested that the introduction of a specific exception to permit TDM “would not negatively impact on the original data provider’s rights and commercial interests because the technology is not intended to reprint the original data, but to provide a synthesised result. These outcomes do not interfere with the economic value of the copyright material nor compete with it.”⁴⁶

Nevertheless traditional publishers remain concerned that an exception for the purposes of text mining would permit and encourage “industrial scale reproduction of content without prior permission of the rights holders ...”⁴⁷ Further, the UK Publishers’ Association argues that an exception could undermine the primary market for journal articles by establishing a means for third parties to ... reconstruct whole articles with a handful of searches.” The Newspaper Society, which represents the interests of Britain’s newspapers, believes that the exception being introduced in the UK has the potential to infringe the Berne 3-step test as it could conflict with the normal exploitation of publishers’ archives.

2.3 Technological challenges

Traditional publishers have raised concerns about the technologies employed in TDM and their ability adequately to service this activity without damage to their normal day to day operations. They argue that customers who have paid to read would experience a significant slowing down of the service available to them and this could result in publishers breaching their contract. Reed Elsevier, for example, believes that 20 researchers crawling their site would significantly reduce its functionality for other users.

Thomson Reuters supports this view, arguing that their system is not configured for third party TDM programmes crawling their systems which is likely to seriously impair if not crash their platforms.⁴⁸ The Royal Society of Chemistry claims that, should the volume of TDM requests rise substantially, it would have to introduce additional server capacity, bandwidth and monitoring to deliver an online ‘on demand’ text mining service.

Researchers reply that these concerns are unwarranted. Dr Cameron Neylon argues that TDM is only a small component of the traffic a public-facing operation should be able to deal with. The Wellcome Trust⁴⁹ believes that the argument put forward by some publishers is difficult to equate with the experiences of open access publishers such as BioMed Central, which already provides access to its entire published outputs without encountering such technical problems.

⁴⁶ AIIA submission to the Australia Law Reform Commission consultation on copyright, 2012

⁴⁷ PA response to the UK Government consultation

⁴⁸ Thomson Reuters response to the UK Government consultation

⁴⁹ The Wellcome Trust is a champion of science, funding research and influencing health policy across the globe.

At a multi-stakeholder workshop organised by LIBER (the Association of Europe Research Libraries) in September 2013 it was noted that publishing infrastructures are already ably dealing with heavy demand from services such as Reddit. Demand for TDM constitutes only a fraction of this. As TDM activities grow they will become a key market differentiator for scholarly publishers and should become part of their core business.⁵⁰

Furthermore researchers argue that publishers have a number of techniques at their disposal for managing programmatic access including:

- Appropriate use of caching to ensure sites can cope with the additional load
- Exclusion rules and “crawl delay” so that robots will not exceed a certain rate
- Running intrusion prevention service to block access to robots that exceed a certain threshold
- Having effective monitoring techniques in place to alert the website manager to load issues
- Using load balancers to delay or throttle excessive traffic

However, Jonathan Clark believes that the publishers’ request that text mining crawlers leave 5 or 10 second delays between successive article downloads, while reasonable, is not scalable. He estimates that a collection of one million articles would take 4-8 months of continuous downloading.⁵¹

2.4 Cultural challenges

Traditionally, authors have assigned their copyright to publishers who, for the most part, built their business models on strictly controlling access as a means of recouping their investment in the upfront publishing costs. With the advent of the digital era these costly functions no longer exist and the value that publishers add to the process has diminished. In today’s digital markets, the most important virtue is effective dissemination – getting content out to those who can use and re-use it. Nevertheless, as Reichman and Okedigi note, publishers have been slow to change - “this web of traditional practices and interests carries into the digital age, even though digital networks offer repeated opportunities to break with the limits of the print model and make whole new dimensions of publishing possible.”⁵²

Further, Reichman and Okedigi believe that “not only have publishers sought to configure the online environment on the model of print media, they have also tried to subordinate the new class of intermediaries that digital technology has generated, the Internet System Providers (ISPs), to their own ends, adding yet another layer of potential barriers and transition costs to the diffusion of research results.”⁵³ Until the

⁵⁰ *The Perfect Swell: defining the ideal conditions for the growth of text and data mining in Europe*. A report from a workshop held at the British Library

⁵¹ Jonathan Clark, *Text Mining and Scholarly Publishing*, PRC, February 2013

⁵² Jerome H Reichman & Ruth L Okedigi, *When copyright law and science collide: empowering digitally integrated research methods on a global scale*, *Minnesota Law Review*, Vol.96, No.4, April 2012, pp 1362-1480

⁵³ *Ibid.*, p. 1463

publishing model changes, the authors argue that funders of scientific research should insist on open access publishing ...⁵⁴

Over the past few years the move to Open Access (OA) publishing has been gaining momentum, supported by many governments and some of the most prestigious universities around the world. For example, in May 2013 the United Nations called for a global drive on open data for development, and an OA policy for UNESCO. By the end of the year, UNESCO had initiated an OA repository. In November 2013 Germany's new ruling Grand Coalition announced a commitment to the legislation, governance and infrastructure – including digitization and repositories – needed for comprehensive OA to publicly-funded research and data.

At the European level, Member States supported the idea of developing broader and more rapid access to scientific publications in order to help researchers and businesses to build on the findings of publicly funded research. In 2012 in a Recommendation to Member States 'on access to and preservation of scientific information'⁵⁵ the European Commission outlined measures to improve access to scientific information produced in Europe. The Commission invited EU governments to define policies so that, in particular, "licensing systems contribute to open access to scientific publications resulting from publicly-funded research in a balanced way, in accordance with and without prejudice to the applicable copyright legislation, and encourage researchers to retain their copyright while granting licences to publishers."

The recommendation complemented a Communication on 'a reinforced European Research area partnership for excellence and growth', which sets out key priorities for completing the European Research Area (ERA), including the optimal circulation, access to and transfer of scientific knowledge. In their late 2013 report, the Expert Group on the 'Recommendations on the Implementation of the ERA Communication'⁵⁶ wrote that "a research-friendly copyright framework is needed to maximise circulation of knowledge" and recommended the Commission "lead the European debate about a research-friendly copyright framework, which assures maximum circulation, access, transfer and re-use of scientific knowledge (with a special emphasis on text and data mining) while protecting the intellectual property rights of authors."

Following on from this the Commission agreed that open access⁵⁷ to scientific publications should be a general principle of the current Horizon 2020 research framework programme. In the model grant agreement for Horizon 2020 the Commission states that the beneficiaries must:

- (a) deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:

⁵⁴ Ibid., p. 1467

⁵⁵ http://ec.europa.eu/research/science-society/document_library/pdf_06/recommendation-access-and-preservation-scientific-information_en.pdf

⁵⁶ http://ec.europa.eu/research/era/pdf/era_progress_report2013/expert-group-support.pdf

⁵⁷ Legally binding definitions of 'open access' and 'access' in this context do not exist, but authoritative definitions of open access can be found in key political declarations on this subject. These definitions describe open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine.

(i) the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;

In all cases, the Commission also encourages authors to retain their copyright and grant adequate licences to publishers. Creative Commons offers useful licensing solutions in this regard (e.g. CC-BY or CC-0 licences).

There is, however, a view that the introduction of variations on the CC-BY licence (e.g. CC-BY-ND) has muddied the waters. While these licences are considered better than previous licences researchers need to be careful about the sub-text and the permissions that are granted through these licences.

Another issue of concern for publishers is attribution. According to JISC “arguably, the key IPR uncertainty in text mining surrounds the inability to attribute every copyright owner/author, due partly to the vast number of articles searched but also because the extent of copying of each article is difficult to audit, and in most – but not all – cases is probably ‘insubstantial’ and may not raise IPR issues, but certainly raises contractual issues.”⁵⁸ Traditional publishers disagree, arguing that while they are willing to support requests for TDM they want to maintain what they see as a basic principle of copyright – that rights owners have a right to prevent anyone using their works without their consent. It is understood that researchers are now able to cite the database rather than each individual author.

Like other industries the publishing industry is being forced to re-examine its business model in light of digital communications technologies and to question whether the current approach is viable in the long-term. At present, the response is to find new ways of licensing largely within the basic model that has existed for some time. Cameron Neylon⁵⁹ is among those who argue that this will not shape a competitive industry in the long-term. “Traditional publishers actions, whether this access initiative, CHORUS, or their grudging approach to open access implementation, consistently focus on retaining absolute control over any potential use of content that *might hypothetically* be a future revenue source. This means each new means of access, each new form of use, needs to be regulated, controlled and licensed. This is perfectly understandable. It is the logical approach for a business model which is focused on monetising a monopoly control over pieces of content. It’s just a really bad way of serving the interests of authors in having their work used, enhanced, and integrated into the wider information commons that the rest of the world uses.”

⁵⁸ JISC response to the Hargreaves review on IP and Growth, 2010

⁵⁹ This is a personal comment, made in an interview with the Expert Review, from Cameron Neylon rather than the view of PLOS.

3. Economic issues

3.1 Basic economic considerations

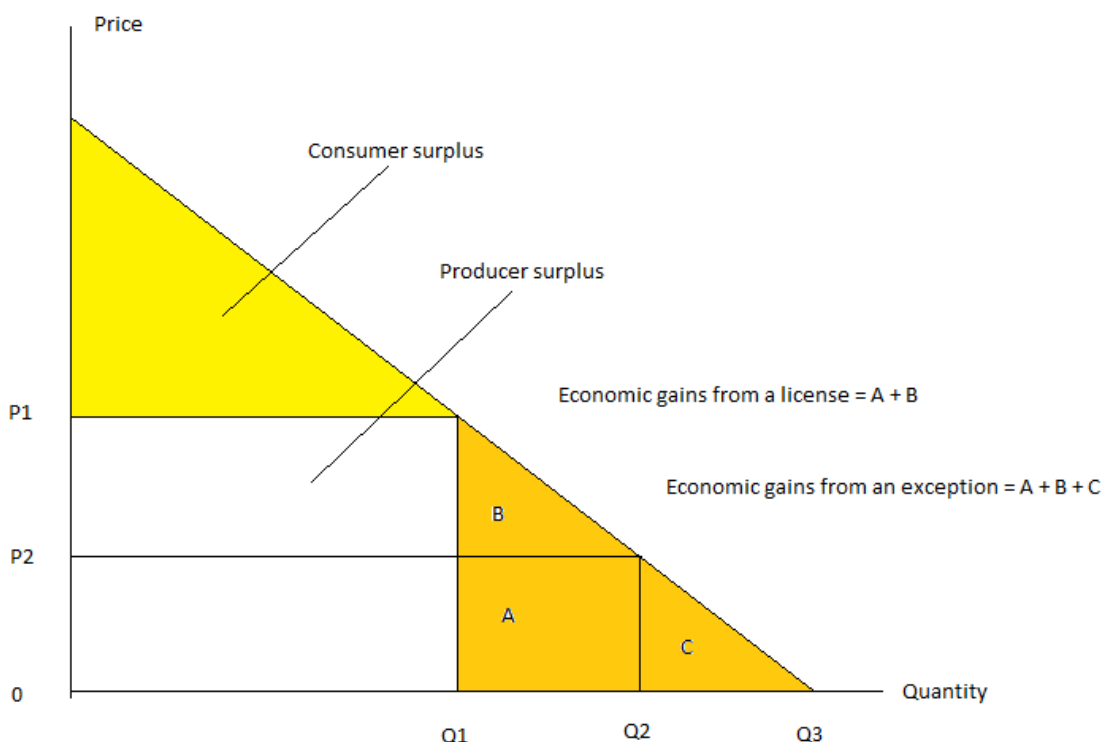
Policy makers should, logically, aim to strike an overall welfare-maximizing balance between the benefits for users and the incentives for copyright holders. This balance is an empirical question; there is no a priori theoretical answer as to what the appropriate degree of copyright protection should be. However, there is hardly any empirical analysis available on the appropriate degree of copyright protection in general, and nothing at all for the case of TDM. The absence of such empirical evidence has resulted in a strongly normative and often antagonistic debate between legal scholars, lobbyists and advocacy groups.⁶⁰

Before we start with empirics, it is important to sketch a very basic economic framework for the analysis of copyright and the impact of possible exceptions. Figure 1 explains the basic economic mechanics of copyright. Copyright attributes a monopoly on the use of an innovative product to a copyright holder. The downward sloping line represents consumer demand for an innovative product: demand is lower when price is higher. The copyright holder can sell the product at a profit-maximizing monopoly price P_1 that leads to the sale of Q_1 units of this product. The white area represents the gains for the copyright holder, the yellow area the consumer welfare surplus (the difference between the price that consumers were willing to pay and the price they actually pay). The orange area is the welfare loss to society: the sales that did not happen as a result of price P_1 . Economists call this area the deadweight welfare loss: all parties lose some potential gains. This is the consequence of giving a monopoly to the copyright owner and the price being fixed above marginal production costs.

Clearly, from this static perspective, copyright is economically inefficient. It is only by adding a dynamic perspective that copyright becomes an economically efficient tool for society: If the copyright owner did not have a monopoly, the price would fall to the marginal cost of making the work available, which may be close to zero in the case of digital information goods. Copyright owners would then have diminished financial incentives to invest in innovation and the supply of innovation would decrease. That would of course reduce welfare for both consumers and producers.

⁶⁰ A notable exception is JISC (2012). This report examines potential research costs savings due to labour productivity gains that TDM may generate (it would speed up data search and analysis). Based on an assumed (but not empirically verified) productivity gain of 2%, it estimates total research cost savings at £127-£158m per year for the UK only. Productivity gains are a source of economic welfare gains. The report does not discuss whether TDM would come in the form of a licensing system or a copyright exception for research. In other words, it omits a key economic factor: do copyright holders receive compensation (and do users pay a price) for TDM or not? Since the JISC report does not discuss potential price savings the implicit assumption in the report seems to be that TDM comes in the form of a copyright exception without compensation. The focus on licensing and consequently on research productivity gains and cost savings is only part of the picture. There is a possible cost side to a copyright exception because it may trigger a negative supply side response in terms of reduced investment incentives for database owners.

FIGURE 1: A simple economic welfare analysis of copyright



Market-based copyright licensing activity produces an output Q_1 at price P_1 . There are non-negligible market failures in the licensing of copyright for TDM, due to transaction costs, externalities and possibly strategic behaviour of rights holders that generate the welfare loss $A-B-C$ (See section on Empirical Evidence below). This may justify regulation that seeks to create legal certainty and a more permissive framework for TDM, for example through a special TDM licensing system that reduces transaction costs or through an exception in law.

How would a more efficient TDM licence or exception affect economic welfare? That is explained in the orange area of Figure 1. A more efficient TDM licence with compensation for the copyright owner would result in a price, say P_2 , to be paid by the TDM user, in return for an additional amount of information (Q_2-Q_1) that can be extracted from the data. The copyright holder would make a profit A , the user would gain a consumer surplus B . There would still be a social welfare loss C for society but the area is much smaller than without a more efficient TDM licence. Clearly, a well-designed licensing system represents an improvement in economic welfare, but the extent of that improvement depends upon the design of the exception and the marketplace response to its terms. Would an exception perform better in economic terms than a licence? As shown in Figure 1: a TDM exception without compensation

for the copyright owner would bring the price down to zero and increase the quantity to Q_3 . All deadweight welfare losses would be eliminated. In this case, an exception would be an economically superior solution provided that the long-term dynamic supply side response would not be significantly negative. A negative impact on the supply of databases for TDM could reduce or even eliminate these welfare gains.

The empirical question is whether total surplus after implementation of an exception would still exceed consumer surplus before the exception. The underlying economic bargain in copyright law is that a positive supply side response over time compensates for the welfare losses of a copyright monopoly. Whether this effect transpires in practice remains an empirical question. Because there is so little empirical research on the efficiency of copyright law, we do not know the answer to this question.

The decisive question, therefore, is how a TDM exception would affect the supply of new copyright works. This question is more easily answered where the production of text and data is publicly financed, intrinsically motivated or where the text and data suitable for TDM is a side-effect of other online activities. It becomes problematic when the supply of works suitable for TDM is very sensitive to licensing income.

For the publicly financed text, data and other media content – for example the output of publicly financed academic research – a copyright exception is more easily justified on economic grounds because public financing is the main incentive to supply work. There is little justification to incur the transaction costs and market failures associated with copyright protection. Subject to appropriate codes of conduct, a copyright exception for TDM of text and data which is already publicly available online could also be justified. The supply of this type of data should not be sensitive to a TDM exception, except where it would adversely affect the accessibility of text and data for other purposes. A compensation system, for instance as in a copyright collecting society, provides an option so long as the transaction costs associated with it seem low and the expected increase in the supply of suitable text and data seems large. This is probably not the case in the main TDM areas discussed in the section below which discusses market failure. In any case, for copyright works that have been created without any incentives for prospective TDM licensing (ie the existing, historical, digital archive) the efficient compensation of rights holders would not exceed the probably modest opportunity costs of making these works available to miners.

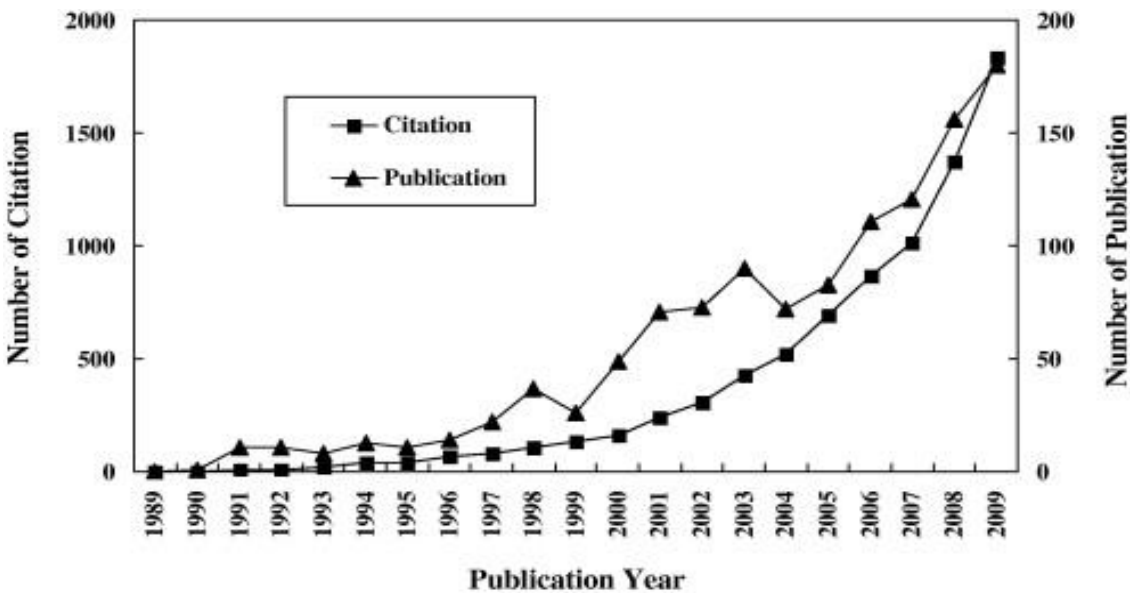
3.2 Empirical evidence

Growth in data mining

There is little publicly available data on the prevalence of TDM. Regarding academic research only, two papers by Tsai (2012; 2013) contain bibliometric data on the diffusion of data mining. Tsai (2012) uses information from the Social Science Citation Index (SSCI) supplied by Thomson Reuters and covering almost 2,000 academic journals in the social sciences and including data from 3,300 leading scientific and technical journals. He recorded the number of academic publications containing “data mining” in topic headers and found 1,181 altogether between 1989 and 2009.

The data assembled by Tsai (2012) shows rapid growth in the number of TDM related publications and their citation counts (see Figure 2). The development conforms well to an exponential growth pattern that is typical for the diffusion of important new technologies. There was sustained and rapid growth over two decades, even though the trend in publications is not perfectly consistent. Despite the difficulty associated with predicting technological change, this would suggest that further rapid growth is very likely.

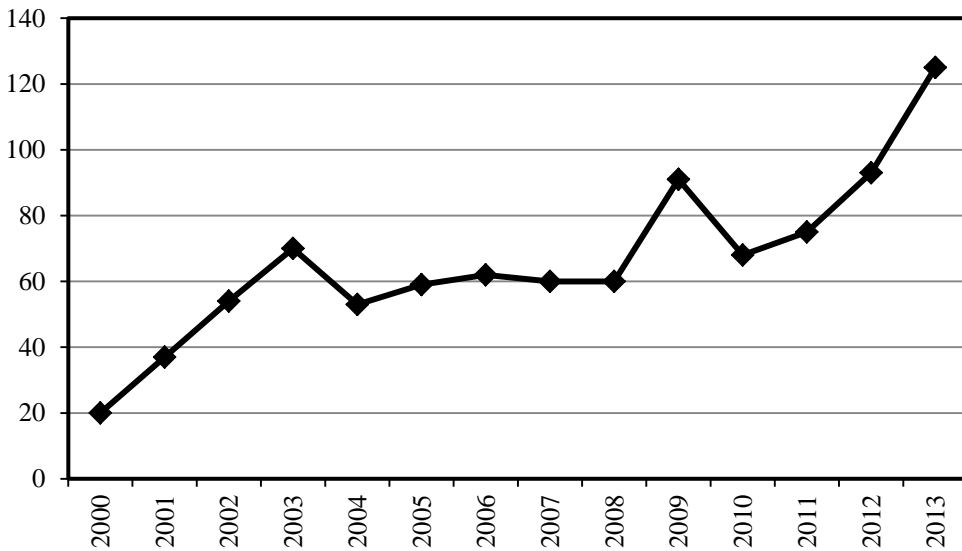
FIGURE 2: TDM related publications and their citation counts



Source: Tsai (2012) based on SSCI.

A forthcoming report by the Lisbon Council examines the number of patents granted in data mining – see Figure 3. This data also shows an upward trend, in particular since 2010, which suggests progress in TDM techniques and expectations of further growth in this area.

FIGURE 3: Patents granted in data mining, 2000-2013



Source: The Lisbon Council (2014) forthcoming.

Finally, a basic exploration of search results (see Appendix) on the search engine Google Scholar demonstrates that TDM accounts for an increasingly large share in total research output. Growth rates over recent years have been high. This outcome is consistent with the secondary data from Thomson Reuters' Web of Science discussed earlier in this section. Data mining related research already makes up a surprisingly large share of publications covered on Google Scholar. Text mining is less frequently referred to in academic work but growing even more rapidly.

The relative performance of European academia in data mining

Tsai (2012) also provides data on the share of TDM-related, academic publications in various countries (Table 1). A rough and ready comparison of this data with SSCI data on publications allows us to consider whether European countries perform similarly in terms of overall research performance and TDM.⁶¹ The data suggests that European countries perform very differently. For example, for Germany, France and Italy the share in TDM publications is less than half that of all publications. The Netherlands and Sweden have similar shares of TDM research output to what would be expected by their overall publication performance. Great Britain has a much greater share of TDM publications than its share in all publications.⁶²

By contrast, the US and Canada account for much greater shares of TDM publications compared to all academic publications. Taiwan – and to some extent Australia – also account for large shares of TDM publications. South Korea has a

⁶¹ The main problem in making this comparison is that Tsai (2012) reports on the overall counts between 1989 and 2012, whereas the available data on countries' share in all publications is for 2010 only. Schmoch et al. (2012, table 3) also contains data on countries' share in all academic publications on SSCI/Web of Science for 2000 to 2010, and these shares are reasonably stable throughout. Since the bulk of TDM related publications are from 2000 to 2009, the main results of the comparison between countries' TDM publications and entire academic publication output are certain to hold.

⁶² Finland performs well in particular regarding citation counts of TDM publications.

similar share for both TDM and other research output. China and Japan publish much less TDM research than would be expected from their overall academic research output.

The mixed performance of European countries in terms of TDM research output may indicate two things:

- Firstly, the British example in TDM research suggests that there is great potential for this type of research in Europe, but that language may be an issue
- Secondly, several large EU Member States such as Germany, France and Italy lag behind the leading countries in this area.

There is scope for more meticulous empirical research to control for intervening factors and to isolate the effect of public policy. It would also be desirable to consider TDM sectors other than academic research.⁶³

⁶³ Tsai (2013) finds that the use of the term “knowledge management” in academic publications has also increased strongly since 1990. Knowledge management is closely related to “data mining”, but typically refers in particular to business management. The concept seems to be well researched in England and Scotland, that together account for 17.66% of worldwide publications on SSCI between 1989 and 2009 (the US accounts for 33.09%).

TABLE 1: Country share of publications with title header “data mining” and citation counts, 1989 and 2009*

Rank in publications (citations)	Country	Number of Publications*	% Share in all publications (1181)**	Citations	Citations per publication
1 (1)	The US	551	46.66	4781	8.68
2 (2)	Great Britain	131	11.09	1159	8.85
3 (5)	Taiwan	104	8.81	436	4.19
4 (3)	Canada	67	5.67	547	8.16
5 (8)	China	54	4.57	187	3.46
6 (6)	Australia	47	3.98	350	7.45
7 (9)	Germany	32	2.71	177	5.53
8 (7)	South Korea	32	2.71	232	7.25
9 (15)	Spain	27	2.29	79	2.93
10 (10)	Netherlands	21	1.78	135	6.43
11 (14)	Belgium	20	1.69	96	4.80
12 (12)	France	20	1.69	105	5.25
13 (19)	Japan	18	1.52	49	2.72
14 (16)	Italy	17	1.44	78	4.59
15 (21)	Brazil	13	1.1	33	2.54
16 (16)	South Africa	13	1.1	69	5.31
17 (22)	Sweden	12	1.02	11	0.92
18 (17)	Turkey	12	1.02	53	4.42
19 (20)	India	11	0.93	30	2.73
20 (23)	Slovenia	11	0.93	4	0.36
21 (21)	Austria	10	0.85	30	3.00
22 (4)	Finland	10	0.85	474	47.40
23 (12)	Singapore	10	0.85	105	10.50

Source: Tsai (2012) based on Web of Science / SSCI.

* Data adds up to 1,243, whereas the column header reports 1.181 publications. This is probably due to double-counting for papers with authors from several countries.

** Shares add up to 105.3%, which is probably because double-counting was not considered when calculating percentages. All values in this column are then biased upwards by ca. one twentieth.

*** For Great Britain, Tsai (2012) separately reported data for England, Scotland and Wales, which are summed up here.

TABLE 2: Countries' share within all SSCI and SCIE publications, 2010

Country	Whole count*	Fractional*
USA	22.2	23.9
China	8.7	9.9
Great Britain	6.2	4.9
Germany	5.8	5.4
Japan	4.8	5.3
France	4.1	3.9
Canada	3.6	3.4
Italy	3.4	3.3
India	2.7	3.1
South Korea	2.6	2.9
Brazil	2.1	2.3
Netherlands	2.1	1.8
Sweden	1.3	1.1
Finland	0.6	0.6
Taiwan	na	na
Australia	na	na
Other countries	28.6	27.3
Total	100	100

Source: Schmoch et al. (2012) based on Web of Science / SSCI (whole counts recalculated).

* Fractional counts include a weighting for publications with authors from several countries.

3.3 Economic consequences of legal reform

In the remainder of this section we attempt to translate the few relevant empirical data points that we have with regard to TDM into a macro-economic impact estimate of reforms to the legal framework governing TDM solutions, either in the form of a copyright exception, without compensation for copyright owners, or as a licence with compensation for copyright owners.

1. We do not have estimates of the market value of all online databases. We only have an estimate of the size of the scientific publishing industry, a very narrow definition of the scientific databases that we discuss here. According to the annual report of the Scientific Technical and Medical publishing Industry Association (STM, 2012)⁶⁴ the size of the worldwide English-language scientific publishing market can be estimated at US \$23.5 billion (2011) or

⁶⁴ See the STM annual report: http://www.stm-assoc.org/2012_12_11_STM_Report_2012.pdf. About 52% of revenues come from the US, 32% from Europe/Middle East, 16% from the rest of the world. Within this overall market for STM information, the global 2011 annual revenues from journals were estimated at \$9.4 billion and those from books (and e-books) at \$3.8 billion. Journals publishing revenues are generated primarily from academic library subscriptions (68-75% of the total revenue), followed by corporate subscriptions (15-17%), advertising (4%), membership fees and personal subscriptions (3%), and various author-side payments (3%).

€18 billion. Slightly less than a third of that is generated in Europe - around €6 billion. We can safely assume that this is essentially expenditure by researchers and research institutions on subscriptions to journals, although part of this expenditure is for educational use such as students' use of journals in university libraries and is therefore not necessarily directly research related.

2. According to Eurostat in 2012 the total research expenditure in the EU27, both public and private, stood at €266.9 billion, which represents about 2 per cent of EU GDP⁶⁵. It has hovered around 2 per cent over the last decade. We can thus estimate the share of scientific publications in total research expenditure at 2.2 per cent.
3. Access to TDM increases the productivity of research because it increases research output with unchanged labour inputs. There are no empirical estimates of the impact of TDM on the productivity effect of research. The UK Jisc study assumed that TDM increases the volume of data accessible to researchers and thereby increases the productivity of research by 2 per cent. If we consider this crucial but unproven assumption to be credible and apply it to EU-wide research expenditure, the real value of research output produced by the EU research budget would increase by 2 per cent or €5.3 billion to total €272.2 billion – for a constant budget.
4. The long-term impact of a change in the volume of R&D production expenditure on GDP has been estimated by various authors. This impact is due to the externalities that research generates in terms of innovative products, productivity and consumer welfare increases. The value of the externalities is usually much larger than the cost of the research expenditure. Here, we take an elasticity estimate by Guellec & Van Pottelsberghe (2004) of 0.13: a 1 per cent increase in R&D expenditure results in a 0.13 per cent increase in GDP. Assuming linearity, a 2 per cent increase in real terms in the research budget would thus result in a 0.26 increase in GDP⁶⁶ or an overall gain of $12500 \times 0.0026 = €32.5$ billion.
5. Note that the estimated elasticity of 0.13 by Guellec & Van Pottelsberghe (2004) is a rather low estimate. In an earlier study (2001) for the OECD the same authors found that the long-term elasticity of government- and university-performed research on total factor productivity is around 0.17. Several other researchers have proposed considerably higher estimates. In an older study, Nonneman & Van Houdt (1996) found that the elasticity of GDP with respect to R&D is twice as high at 0.23. Archaya and Coulombe (2005) found a value of 0.24 to 0.50, two to four times higher. Our estimate of a €32.5 billion gain could thus be considered as a lower limit, given a research productivity increase of 2 per cent. Even if the average increase in research productivity as a result of TDM were to prove much lower than assumed by

⁶⁵ See http://epp.eurostat.ec.europa.eu/portal/page/portal/science_technology_innovation/data/database

⁶⁶ EU GDP in 2012 is estimated at 12.5 trillion Euros.

the Jisc (2012) study, a GDP gain with an order of magnitude of tens of billions of Euros would still be feasible.

6. Moreover, the above estimate is limited to the narrowest TDM definition - the market for published scientific research only. Extending the TDM definition to a wider market would significantly amplify the economic impact, though there is no data on which to estimate the the scale of this.
7. In the short-run, remuneration of publishers from the research budget involves only a static shift in welfare between two groups in society. It may however affect social welfare because remuneration systems are costly to operate. They require an organisation to operate the compensation of the copyright holder (e.g. a collecting society) and would entail negotiations, monitoring and enforcement. These would entail transaction costs and deadweight losses for society.
8. In the long run, when we include the effect of TDM on the supply of input works, the situation may be different. The externalities generated by an increase in research output produce additional welfare gains for consumers and producers of copyrighted content. The decisive question is whether compensation of rights holders for TDM is necessary to sustain the supply of suitable input works open to TDM by researchers. Compensation systems may also encourage rights holders to develop supporting services to facilitate TDM by rights holders. Remuneration for additional services offered by rights holders can of course exist in parallel with a TDM exception.
9. The social benefits of additional compensation for rights holders for TDM uses are probably lower than the costs of running a compensation system in the following cases:
 - (a) Where existing works are concerned, so that only the costs of making works available and developing support services for TDM by rights holders are concerned. (With a TDM exception and greater legal certainty, users would have greater incentives to develop new solutions in this area.)
 - (b) Where the supply of new input works is mostly financed through other means, for example public finances in the case of most European academic research.
 - (c) Where intermediaries enjoy extensive market power so that they may appropriate an excessive share of licensing revenues and make super-normal profits (rather than passing on revenues to creators of input works or financing efficient amounts of innovation in intermediary services).
10. The analysis so far is based only on research productivity gains and the implied gains in research output (for a constant research budget). Even if a TDM licensing system would compensate and entirely transfer the productivity gains from researchers to publishers, there may still be other potential sources of gains in research quantity and quality from TDM. TDM may enable the emergence of new research projects that were simply not possible before

digital TDM technology. As such, TDM could shift research expenditures to different types of projects. TDM may also increase the quality, accuracy and reliability of existing research projects because it allows access to a much wider dataset. We have no information on these potential gains and therefore cannot provide any empirical estimates.

11. There are countervailing effects of a copyright exception for TDM (or a compensation system regulated to charge low user fees). On the one hand, an exception could displace demand for private licences of copyright works. On the other hand, TDM increases the productivity of research – and thus the total economic value of research output – so that demand for related services will increase. Put simply, the results of TDM research will also be published.
12. As always, the economic effects of copyright protection involve an empirical question and depends on the balance between the short-term static equilibrium (the level of copyright protection, in this case the additional remuneration accorded to the copyright holder for TDM) and the long run dynamic equilibrium (the labour productivity gains and quality gains for the users of TDM and the ensuing increase in GDP). There is no a priori theoretical answer to these questions and therefore no precise figure which can be attached to the scale of the welfare benefits attaching to variations in the licensing or legal regime.

3.4 Market failure: what prevents competitive TDM in Europe?

According to our estimates, the stakes in TDM related research are high and large parts of the European Union are lagging behind the most successful countries in this area. This section discusses potential market failure regarding copyright and transformative use of copyright works, which some legal scholarship and jurisprudence suggest is the correct way to view the outputs of TDM.

The economic literature identifies three fundamental reasons why the transformative use of copyright works – creating new valuable works by building on preceding works – may not approximate a socially efficient level with effective copyright protection: (a) transaction costs, (b) strategic behaviour by copyright holders and (c) externalities. Some other arguments have been added, though they can usually be presented as special cases of (a) to (c). We focus here on the three main arguments.

Transaction costs

The debate on TDM has been mainly confined to legal scholars and the law and economics literature. Traditionally, the latter follows a Coasian transaction costs approach to copyright and to copyright exceptions. Copyright law is usually presented as necessary to overcome a market failure to deliver a sufficient production of innovative artwork like music, films, books, newspaper articles, etc. Since artwork is non-rival and hard to make excludable, producers would not have a sufficient financial incentive to produce the artwork, because once produced it would be available to all at a very low reproduction cost. From a Coasian perspective, for an artwork to be produced in the absence of copyright law would require costly

direct bargaining between producers and consumers. These are transaction costs. Since they would be high compared to the value of the product, they would have the effect of depressing the supply of artwork.

From a Coasian perspective, copyright law is a device that reduces transaction costs and so facilitates the production of artwork. In the absence of transaction costs, copyright-protected databases will be traded and used efficiently, irrespective of who holds the rights initially. Copyright owners with market power may price-discriminate against others so that all the welfare benefits accrue to them, but from a societal point of view this would still be welfare maximizing. As a corollary, in the presence of transaction costs, for instance costs related to negotiating a deal with many copyright owners, a welfare-enhancing agreement is not assured. In that case, the purpose of an efficient TDM licensing system would be to diminish transaction costs. It would still result in a compensation for copyright owners.

The argument can be extended to copyright exceptions. Exceptions limit the scope (coverage) of copyright and are economically justified when transaction costs are so high that they would prevent a copyright transaction from taking place. If no efficient and transaction-cost-reducing TDM licensing system can be designed then it would be better to legalize unauthorized use by means of a TDM exception. Without an exception, in these circumstances, TDM would either not occur or would occur on a significantly diminished scale, thereby generating “deadweight loss” for society: welfare losses that benefit neither the producer nor the consumer.

On the other hand, if a low-cost and efficient TDM for research licensing system could be designed there would be no need for an exception since the market would be able to deliver licences at low transaction costs and thereby enable transactions to take place. In theory, TDM licensing would involve low transaction costs if it involves only one copyright holder, say a single journal publisher or database owner, and one user. The two parties could negotiate a deal directly.

It is often argued that transaction costs in the market for copyright works would fall with digitization, making the market more efficient (e.g. Depoorter and Parisi, 2002). Production costs to bring large datasets online, search costs to identify a suitable data source for TDM and search costs inside these large databases have indeed fallen online. However, bargaining and contracting costs have probably not decreased substantially. The contrast between the dramatic drop in digital information costs and the still high transaction costs for (mostly analogue) bargaining are at the source of the current TDM debate. What is more, total transaction costs in a market are a function of the number of transactions and the costs per transaction. With lower search costs and lower costs of accessing works online (with or without authorisation from rights holders), users have diversified their consumption, which increases the number of potential transactions and could thus increase total transaction costs.

Strategic behaviour

Researchers (Gordon & Bone 1997, Depoorter & Parisi 2002, Lemley & Shapiro 2002) have pointed out that this transaction cost approach has its limitations and that there may be several other reasons to limit the scope of copyright and grant exceptions without compensation.

Strategic behaviour by copyright holders may drive up the price (not the transaction cost) of licences. This phenomenon has been extensively studied and documented in the case of patents. Lemley and Shapiro (2002) point out that the patent system was designed with a paradigmatic invention in mind - a single innovative product covered by a single patent. In reality, innovative products are becoming more complex and contain increasingly large numbers of patents. The stacking of patents in a single product makes royalty negotiations more difficult. The authors refer to mobile phones as an example of a patent thicket that may well include thousands of patents. A single patent holder could hold-up the entire production of a new phone and demand unreasonable compensation. They develop a game theoretic framework to show how this may lead to royalty charges far above a "fair" monopolistic price.

A similar point can be made for copyright. It was designed with a single copyright-protected expression of creativity in mind. In reality, creativity can be cumulative and innovative artwork can build on prior copyright-protected products. Prior copyright holders who are able to price discriminate against downstream innovators may actually charge prices above a monopolistic rate if 'hold-up' problems occur. The hold-up problem is well known in the transaction cost literature (Williamson, 1985) but there are no obvious market-based solutions for this problem since contracts are always incomplete. Depoorter & Parisi (2002) follow a similar line of reasoning but apply it directly to copyright. Not only do transaction costs account for the "tragedy of the anti-commons", strategic behaviour by copyright holders may prevent some transactions from materializing. In the same vein as Lemley & Shapiro they argue that multiple copyright holders of complementary (non-substitutable) inputs into an innovative product can result in substantial deadweight loss of unproduced innovation because profit maximizing copyright holders will push up the price of licences. Full substitution would eliminate this deadweight loss. However, since copyright holders operate almost by definition in a monopolistic competition market, full substitutability is unlikely to be the case.

Even in the absence of strategic overpricing behaviour, the monopoly granted to a copyright holder will only result in maximised social welfare if all users who are willing to pay at least the marginal cost of reproducing the copyright-protected content are served. This implies that the copyright holder is able to practise perfect price discrimination and modulate the pricing of the copyright licence in such a way that it adapts to the purchasing power and value of the product for each potential user. It is possible to devise partial price discrimination solutions, such as different pricing levels and metering of use, but they remain inevitably partial. If not, deadweight losses will occur and overall social welfare will be reduced as a result of a TDM licensing system. It is not difficult to see why perfect price discrimination behaviour is unlikely to occur. Like music, film and book sellers, database sellers usually offer fixed prices, with limited flexibility. They fix their prices at an assumed profit-maximizing level. That is why the copyright system almost inevitably generates deadweight losses.

Even in the (infrequent) case of a TDM research activity involving only one copyright holder and one user and so with low transaction costs, the hold-up problem can occur. The copyright holder may simply not be interested in negotiating a TDM deal with a researcher because the copyright holder's main source of revenue may not be related to research. This is the case for many datasets that are publicly available

and accessible on the internet but that explicitly exclude data mining, which is beyond the margin of the rights holder's core business model.

This could also be the case where valuable information is rival in use. A researcher who enjoys exclusive access to valuable data has an advantage over competitors and so do firms that have exclusive information on market conditions. The individual utility derived from data will then decrease with the number of other relevant users. This may result in a coordination problem (a prisoner's dilemma), where individual rational behaviour does not result in the best outcome for society at large. The reason is that each supplier of data will want to avoid a situation where he makes 'his' data available to others who do not respond in kind. If nobody has an incentive to move first, the benefits of TDM may not be fully realised. Public policy could seek to break such an inefficient equilibrium by setting a universally adopted standard in which suppliers of data mutually make their data available to each other: in effect a publicly mandated and funded 'commons'.

Externalities

TDM is likely to generate positive externalities similar to the externalities associated with research spending in general. The outcome of research may increase productivity for a large number of agents and firms, and stimulate GDP growth, thereby benefiting many people. These benefits are not accounted for in the negotiations between a copyright holder and a researcher. The bargaining done is a function of the copyright holder's private benefits and the researcher's research budget. The spill-over effects on other people's welfare are not accounted for. Externalities drive a wedge between the private and the social value of a transaction. As a result, the number of transactions that materializes is lower than the socially optimal number.

From a Coasian transaction cost perspective, these externalities may be internalised provided that the transaction costs associated with doing this are fairly limited, compared to the value of the deal. It is easy to see that this is unlikely to be the case for the spill-over effects from research: how to involve all the (potential) beneficiaries of TDM for medical research, for instance, in a negotiation with the copyright holder on accessing a medical database?

Information in general and digital data in particular are not depleted through use. They tend to be non-excludable so that they can generate external benefits. With incomplete information on potential users of the data, rights holders cannot price discriminate accurately. The result is that copyright holders are not able to appropriate all of the value of the works to which they hold the rights. They will maximise their private returns without consideration of the wider social benefits and externalities.

A more transaction cost efficient solution is possible in the case of government-funded research: all taxpayers contribute to the cost of the research in proportion to their income and expenditure and so it is logical to assume that TDM access is permissive. Similarly, it could be argued that a government-sponsored scheme might be initiated to finance TDM licences. Similar systems exist in some EU Member States, for instance in the form of additional taxes on digital information

storage hardware (such as USB sticks) to compensate copyright holders for loss of revenue from private copying.

3.5 The scope for special copyright arrangements for TDM

The economic justification for public investment in copyright protection is that without copyright the supply of creative works would fall much below its socially desirable level. The extent to which this problem exists in practice depends on specific market conditions.

We can distinguish several categories of databases to which TDM could apply, starting with the broadest:

1. XXL definition: all databases behind a firewall (as distinct from a paywall). That includes companies' and organisations' internal databases that are not accessible to the public. They require passwords, security clearance and other authorisation for access. We exclude this category from further consideration for TDM because we consider that TDM is not meant to confiscate data that are not in the public domain. Excluding this type of data also potentially resolves the security and privacy issues that may arise. If databases have privacy issues, they should not be in the public domain at all, e.g. health and financial transactions databases. If researchers are seeking access to such databases they should negotiate this directly with the owners on a case-by-case basis. At most, guidelines on good practice could be developed.
2. XL definition: all publicly accessible databases not behind a firewall or a paywall. This data is already in the public domain and can be accessed and observed by anybody at zero-price, e.g. the freely accessible parts of newspaper websites, product and services information available on e-commerce websites, on airline and other transport sites. A TDM exception, without compensation, would not have any impact on the revenue of the owner since the underlying business model does not depend on selling these data; they are already available free of charge. Reproduction of the data for the purpose of other commercial activities may however create competition between the original owner and a new owner that may affect the revenue of the first. Re-publication of the input data for commercial use should therefore, arguably be excluded.
3. L definition: all publicly accessible databases located behind a paywall. Anybody willing to pay the access price can see the data, e.g. the subscription part of online newspapers. A TDM exception would not change the revenue stream for the copyright holder. In the case of newspapers, normal revenue comes from subscriptions that users pay for their daily news reading, along with other revenue sources such as advertising. Researchers are presumably not interested in reading the content of the newspaper articles for their own direct consumption but only in order to derive or aggregate findings in a way that does not substitute for selling news. The risk of a financial disincentive for investment should thus be very small. A TDM licence with compensation would probably bring additional windfall profits for the copyright owner, over and above the revenue already generated by "normal" (non-TDM) use.

4. M definition: all publicly accessible databases behind a paywall whose clients are mainly researchers and whose revenue stream is derived mainly from research expenditures by private and public organisations, e.g. Reuters, Bloomberg, Nielsen, ComCast, GfK. Substitution risks exist if the output produced by the researcher competes directly with the normal stream of outputs produced by the copyright owner. If the normal stream of revenue is derived from selling primary or input data and if the TDM exception prohibits re-publication of the primary data, then the substitution risk is marginal⁶⁷. If the normal revenue stream comprises copyright owners' own research output, then substitution risks are higher. For example, if a researcher produces an economic study with aggregated data from Bloomberg, such a report may compete in the market with Bloomberg's own reports. For this reason, database owners sometimes include clauses in a user agreement that prohibit the publication of competing products. Nevertheless, the variety of reports that can be produced using these databases is so wide that direct competition in this very heterogeneous market for research reports is likely to be small. For this reason, most of these copyright holders allow the use of their data for research purposes and actively sell their databases to the research community.
5. S definition: scientific publishers' databases behind a paywall, e.g. Elsevier, Springer, etc. This was the core issue under discussion in the Licences for Europe working group on TDM. Again, the question is whether the TDM research output would be a substitute for the normal revenue stream generated by the primary data produced and sold by the publisher. Scientific publishers are generally not in the business of producing research reports themselves. A TDM exception would therefore not diminish their normal revenue stream. Publishers prefer TDM licences because it gives them an additional (windfall) source of revenue.

The potential risk of a negative supply side response and risks from revenue substitution between the original data and the TDM data output go hand in hand. This is where it is crucial for copyright policy to define appropriately the scale and scope of any special arrangements made to facilitate TDM.

TDM seeks to extract new information or new insights from existing digital data; insights that could not be readily observed in the existing data without a computational effort. This transformative use needs to be distinguished from reproductive use that simply reproduces the original data. Reproduction is usually an essential first step in TDM research. The decisive issue is that TDM researchers also incur development costs for creating information goods and services. By definition, the output of a TDM process contains a different information set than the information provided by the rights holders of the original and probably diversely owned datasets.

Without entering into legal considerations in this section of the report, the above definition of TDM has important implications for the economic analysis that we focus

⁶⁷ The risk of straightforward piracy always exists, even with normal paywall access. This risk cannot be attributed to a TDM exception. Even without a TDM license or exception pirates can always scrape an entire database.

on here. The task for policy makers is to identify situations where the incentivisation of more extensive TDM research does not adversely affect the supply of input works.

TABLE 3: Domains for TDM and substitution risks

Domains for TDM and substitution risks	
Type of datasets	Revenue source
XXL - datasets behind a firewall, not in the public domain	Excluded from TDM
XL - all publicly available datasets not behind a firewall or paywall	Revenue, if any, derived from other commercial uses
L - publicly available datasets behind a paywall	The paywall provides sufficient revenue from other sources
M - publicly available datasets behind a paywall used mainly for research purposes	Paywall provides - unless TDM research substitutes for own output
S - scientific publishers' datasets only	Paywall provides sufficient revenue. Publishers do not produce research output, so no substitution

3.6 An exception for TDM for non-commercial research only

This brings us to a related issue - whether to restrict a TDM licence or exception to non-commercial research only or to allow it for all types of research. Here we do not enter into the legal debate on the meaning of that distinction⁶⁸ but instead limit ourselves to economic arguments. Our conclusion is that from an economic perspective, making a distinction between commercial and non-commercial use is not very meaningful.

⁶⁸ There seems to be no jurisprudence on the 'non-commercial' character of research though it is mentioned in the EU copyright Acquis. The Explanatory Memorandum to the Information Society Directive reveals that the intention of the legislator is to consider the 'commercial' character of an activity rather than of the 'institution' carrying out this activity. This is a rather vague and arbitrary separation that creates a lot of uncertainty for researchers. What is important however is that data mined through TDM would not displace commercial sales for the original input data owners. With the requirement that the input and output data set are different in content, there can be no displacement.

First, the potential risk with 'commercial' research does not reside in the legal status or private motives of the researchers or their organisation. It resides in the potential risk of sales displacement for the original copyright owner: it is an economic risk. Excluding research by private companies is not a good criterion on which to gauge or reduce that economic risk. Academic research may also lead to the development of commercial products at a later stage. For example, much university research in bio-medical, genetic and natural science may result in commercial products. University research necessarily rivals and competes with privately-financed research. However, that does not imply that the output of private or publicly financed TDM would substitute for the revenue that copyright holders derive from the data on which TDM was carried out.

Second and more importantly, both commercial and non-commercial research can be welfare enhancing for society and should therefore be stimulated by the IPR regime. Indeed, the principal economic argument that we advanced earlier in this chapter in favour of a TDM exception revolves around the externalities produced by research output in general, irrespective of the legal or commercial status of that research. The long-run impact of an increase in the volume of research on GDP can be estimated separately for publicly and privately-financed research but the elasticity coefficients are not very different. If this externality argument is accepted as the primary economic argument in support of a TDM exception, then there is no economic argument to support a distinction between private and publicly-financed TDM.

A well designed copyright regime should provide appropriate stimulus for all types of research and at the same time an appropriate level of protection for all rights owners. Once this balance has been reached, there is no reason to distinguish between commercial and non-commercial research. The database owner should be protected from practices that negatively affect their revenue, not from practices that do not affect that revenue. Even this statement needs qualification - the database owner should be protected against practices that negatively affect revenue in so far as it would reduce overall social welfare. In some cases, negative revenue effects may be more than compensated for by welfare benefits.

4. Legal issues

On the basis of issues raised in the previous sections of this report, the question that this section seeks to address is whether legal barriers impede the conduct of text and data mining (TDM) of databases for research purposes and if so, how these barriers could best be alleviated in the light of the current European legal framework, taking the interests of all stakeholders concerned into account.

Before going into the European situation, it is appropriate to examine how Europe's main trading partners deal with TDM issues in their intellectual property regimes. To this end, this chapter briefly considers the copyright laws of the United States, Australia, Canada, Israel and Japan to see whether TDM activities are permitted and if so, on what grounds and under what conditions. Taking a descriptive approach, the chapter goes on to provide an overview of how databases containing all sorts of works and information are protected under existing European intellectual property law and how the law could support TDM activities for research purposes. The rules laid down in the European Database Directive and the Information Society Directive, as interpreted by the Court of Justice of the European Union and legal commentators, is considered. It focuses essentially on the scope of protection granted to rights owners under the copyright and *sui generis* database regimes and on the exceptions that have been recognised for the benefit of research.⁶⁹

The chapter then sets out a normative approach to consider how the copyright and *sui generis* database regimes could be adapted to permit certain acts of TDM. This could be achieved in several ways, either through an adjustment of licensing practices, through a revised normative interpretation of the 'reproduction right', or through the introduction of an exception on copyright and the *sui generis* database right. Should an exception be introduced in the European legal framework, the question would arise as to whether it should be open to over-riding through the enforcement of restrictive contractual clauses or technological protection measures.

This chapter contains two additional subsections aimed at providing a complete view of all legal issues relevant to TDM activities. The first concerns the unresolved issue of the database providers' power to prevent access and block the use of non-IP protected databases by relying purely on contracts and technological protection measures. The rules on competition may here provide some relief but only in certain specific circumstances. The second subsection highlights the most pressing issues bearing upon TDM from a data protection perspective.

For the purposes of this chapter, TDM is understood to occur through the use of 'digital mining techniques to process huge amounts of texts or data'.⁷⁰ The emphasis is therefore put on the use, in bulk, of the content of compilations or of databases containing data, works, or other subject matter, rather than on such individual

⁶⁹ Generally, see: J.-P. Triaille, S. Dusollier, et al., *Study on the application of Directive 2001/29/EC on copyright and related rights in the information society*, De Wolf and partners, PN/2009-35/D, Brussels, December 2013.

⁷⁰ Ibid., p. 355.

works, data or other subject matter. This distinction is important insofar as the scope of intellectual property protection varies if one considers only the database or also its content, as the object of protection.

4.1 TDM outside Europe

How do Europe's main trading partners deal with the issue of TDM in their intellectual property laws? Are TDM activities permitted without the prior authorization of the rights holder in the United States, Australia, Canada, Israel or Japan? Are researchers in these countries confronted with legal barriers that prevent them from engaging in TDM activities? It is important to note at the outset, that none of the countries examined below have enacted an intellectual property regime that is comparable to the European Database Directive. Among the countries studied here, only Japan offers extra protection against the misappropriation of databases by competitors. The legal regime relevant for TDM activities outside Europe is copyright law.

United States

TDM was considered a relevant factor in assessing whether the Google Books programme would fall within the scope of the 'fair use' defence. The 'fair use' doctrine was developed by US courts and codified in § 107 of the US Copyright Act 1976.⁷¹ The fair use defence is characterised by the open-ended list of purposes for which the use of a work may be regarded as fair, marked by the words 'such as', and by the four factors to be considered in determining whether or not a particular use is fair. The Google Books programme consists of two programmes: the "Partner Programme" involving the hosting and display of material provided by book publishers or other rights holders, and the "Library Program" involving the digital scanning of books in the collections of several public and university libraries. These programmes entailed several activities including making text available and offering tools for online searching of the content of the books and displaying "snippets" of the books.

After the rejection of the proposed settlement between The Authors Guild and Google in March 2011, The Authors Guild continued its lawsuit against Google and at the same time sued HathiTrust, a partnership of major academic research libraries that relies on Google Books Search to create a digital archive of library materials (the HathiTrust Digital Library, or "HDL"). Works within the HDL are used for three purposes: (1) full-text searches; (2) preservation; and (3) to facilitate access for print-disabled persons. In both cases, the Federal District Court of New York had to

⁷¹ US Copyright Act 1976, § 107 reads: 'the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. In determining whether the use made of a work in any particular case is a fair use the factors to be considered shall include: the purpose and character of the use, including whether such use is of a commercial nature or is for non-profit educational purposes; the nature of the copyrighted work; the amount and substantiality of the portion used in relation to the copyrighted work as a whole; and the effect of the use upon the potential market for or value of the copyrighted work. The fact that a work is unpublished shall not itself bar a finding of fair use if such finding is made upon consideration of all the above factors.'

rule whether digitisation of books is a legally fair use of copyright material. The decisions were rendered by different judges (on October 10, 2012⁷² and November 14, 2013⁷³ respectively), both of whom ruled against the Authors Guild and in favour of the application of the fair use doctrine.

Considering the different goals of the Mass Digitization Project under the first fair-use factor, the Court stressed that these were to be considered as transformative uses, referring – amongst others – to the new areas and methods of research, such as **text mining**, that these digital copies enabled. Although one might have expected Google's fair use defence to be weaker than the libraries', Judge Chin in *Authors Guild v. Google* equally affirmed that Google's use of the copyright works in the context of its book scanning and indexing project constitutes "fair use" under copyright law. The court held that Google's digitisation of books is "highly transformative," adds value, serves several important educational purposes, and may enhance the sale of books to the benefit of copyright owners. Again, the fact that Google Books facilitates search, offering an important tool for readers, scholars, researchers, libraries and others to identify and find books, and opens up new fields of research, in particular through **text mining**, was put forward to demonstrate the transformative character of Google's use of the copyright works. In *Authors Guild v. HathiTrust*, the Court refers in a footnote to **text mining** as "new areas of non-expressive computational and statistical research". Admittedly, the Court did not address *as such* any intermediate copying activities by TDM researchers themselves. However, considering the outcome of both *Authors Guild v. HathiTrust* and *Authors Guild v. Google* – concluding that HathiTrust and Google's use of the copyright works met all the legal requirements for fair use – it seems reasonable to assume that copying acts by TDM researchers for the purpose of extracting non-expressive metadata, could be considered fair use under US law.

Canada

The Canadian Copyright Act has contained a fair dealing exception since its initial adoption in 1911. To be exempted under the fair dealing exception, the purpose of the dealing must qualify as one of the allowable purposes under the *Copyright Act*, namely research, private study, criticism, review or news reporting. Secondly, the dealing must be fair. Whereas the Canadian fair dealing exception traditionally received a narrow interpretation compared to the US fair use defence, recent jurisprudence from the Supreme Court of Canada has broadened its scope significantly. In a landmark case⁷⁴, the Canadian Supreme Court was asked to decide upon the application of the fair dealing defence for purposes of research and private study. In the *CCH* case, the Court ruled that 'these allowable purposes should not be given a restrictive interpretation or this could result in the undue restriction of users' rights' (para. 54). The Court added that 'in assessing the character of a dealing courts must examine how the works were dealt with. If multiple copies of works are being widely distributed, this will tend to be unfair. If,

⁷² Text of the decision available from: <http://docs.justia.com/cases/federal/district-courts/new-york/nysdce/1:2011cv06351/384619/156> It should be noted that The Authors Guild has appealed both the decision in *Authors Guild v. HathiTrust* and the ruling in *Authors Guild v. Google*.

⁷³ Text of the decision available from: <http://www.nysd.uscourts.gov/cases/show.php?db=special&id=115>

⁷⁴ *CCH Canadian Ltd. v Law Society of Upper Canada*, 2004 SCC 13 at para 48, [2004] 1 SCR 339 <http://scc.lexum.org/decisia-scc-csc/scc-csc/scc-csc/en/item/2125/index.do>.

however, a single copy of a work is used for a specific legitimate purpose, then it may be easier to conclude that it was a fair dealing. If the copy of the work is destroyed after it is used for its specific intended purpose, this may also favour a finding of fairness' (para. 55). The Court in *CCH* also stated that the allowable purposes must be given a "large and liberal interpretation", and that "research" is not limited to non-commercial or private contexts (para. 51).

The Canadian Copyright Act was modernized in 2012 with, among other important modifications, the introduction of an exception for fair dealing for the purpose of education. This, together with the very broad interpretation given by the Supreme Court to the fair dealing provision in five decisions rendered in 2012, makes the Canadian fair dealing exception almost comparable to the US fair use doctrine.⁷⁵ Considering the Supreme Court's twice reiterated opinion on the importance of allowing fair dealings for purposes of research and private study, it could be argued that TDM activities would probably qualify as a fair dealing under the new Canadian copyright regime.

Australia

Like Canada, the Australian Copyright Act allows fair dealings of works for specific purposes. Unlike Canada, however, the Australian fair dealing exception has not received such a broad interpretation from the courts. As the Australian Law Reform Commission points out, 'where the data mining process involves the copying, digitisation, or reformatting of copyright materials without permission, it may give rise to copyright infringement' under current law. It is unclear whether data mining, if done for the purposes of research or study would be covered by the fair dealing exception, if the whole dataset needs to be copied and converted into a suitable format. Such copying would be more than a 'reasonable portion' of the work concerned.⁷⁶

Israël

The 2007 Act shifted Israeli copyright law from a British 'fair dealing' framework to an American 'fair use' framework, accompanied by an additional list of exceptions. The 'fair dealing' defence is in principle much narrower than the US inspired 'fair use' defence. The main difference lies in the fact that the purposes for which the defence is admissible are enumerated exhaustively in the act.⁷⁷ Fair dealing is therefore not an open norm and the interpretation of the purposes listed in article 2(1)(i) of the former Act by the Israeli courts gave rise to some tension in the years preceding the copyright reform.

Since the amendments of 2007, the Israeli Copyright Act contains an open-ended fair use defence that can be invoked in a wide variety of cases and situations. Article 19 of the Copyright Act of 2007 is modelled after section 107 of the US Copyright Act of 1976 but contains an interesting feature in paragraph (c) according to which

⁷⁵ Michael Geist, *Fairness Found: How Canada Quietly Shifted from Fair Dealing to Fair Use*, in M. Geist (ed.), *The Copyright Pentology*, Ottawa, University of Ottawa Press, 2013, pp. 157-186.

⁷⁶ See : Australian Law Reform Commission's analysis at <http://www.alrc.gov.au/publications/8-non-consumptive-use/text-and-data-mining>

⁷⁷ Meera Nair, 'Canada and Israel – Cultivating Fairness of Use', PIJIP Research Paper, No. 2012-04 American University, Washington College of Law.

the Minister may make regulations prescribing conditions under which a use shall be deemed fair. The amendments of 2007 were not only limited to the implementation of the fair use defence. An extensive number of additional exceptions were introduced in the Israeli Copyright Act covering a number of different uses of works, none of which are directly applicable to TDM activities. The new Israeli fair use provision has yet to be tested in a TDM case. At this time, it is impossible to predict how a judge would rule on the issue, but it is fair to say that in rendering judgment in new situations Israeli courts tend to look to American case law.

Japan

In 2009 Japan introduced, alongside other limitations, an exception aimed at boosting the country's internet economy,⁷⁸ an exception specifically designed to permit TDM. The Japan Copyright Act (2011)⁷⁹ contains an explicit provision to allow text mining, where Article 47^{septies} reads:

'For the purpose of information analysis ('information analysis' means to extract information, concerned with languages, sounds, images or other elements constituting such information, from many works or other much information, and to make a comparison, a classification or other statistical analysis of such information; the same shall apply hereinafter in this Article) by using a computer, it shall be permissible to make recording on a memory, or to make adaptation (including a recording of a derivative work created by such adaptation), of a work, to the extent deemed necessary. However, an exception is made of database works which are made for the use by a person who makes an information analysis.'

A report issued by the subdivision on Copyright of the Council for Cultural Affairs in January 2009 presents the following examples of information analysis: (1) website information analysis and language analysis in which the use of a specific language or character string is analysed and statistically processed and (2) sound analysis and video/image analysis in which the meaning of the sound wave, video, character string, etc., comprising a certain sound, video, image, etc., is analyzed. Although the types of works subject to this provision are not limited, the reverse engineering of computer programming falls outside the scope of this exception: reverse engineering cannot be regarded as "information analysis" because no statistical analysis is conducted.

The rather obscure wording of the last sentence of the provision may be due to difficulties in translation. According to the AIPPI⁸⁰ report of the Japanese Group, when the results of information analysis are presented, it is prohibited to exploit the works subject to the information analysis. The results may be presented or provided only if the results are presented or provided in the form of statistical data, etc., in which the works subject to the analysis are not exploited. Recently, Japan has seen the introduction of new services that enable users to search and analyse users' comments on the Internet including blogs, review sites and social media. The

⁷⁸ Yoshiyuki Tamura, Rethinking Copyright Institution for the Digital Age, 1 W.I.P.O.J. 63-74 (2009)

⁷⁹ Japan Copyright Act: <http://www.cric.or.jp/english/clj/cl2.html>

⁸⁰ The AIPPI is The International Association for the Protection of Intellectual Property.

establishment of the said Article is one of the factors that have promoted the emergence of those new services.⁸¹

4.2 TDM and European Intellectual property protection

Scope of protection

Whereas scientific publications virtually always attract copyright protection under the copyright laws of the Member States of the European Union, compilations of data, works, or other subject matter may not so easily fall under the copyright regime.⁸² Since copyright does not protect mere facts and ideas, but rather attaches to the original expression of ideas, compilations of data, works, or other subject matter may not easily qualify as protectable subject matter due to a lack of originality. The concept of originality in copyright law has been harmonized at the European level with respect to software,⁸³ databases⁸⁴ and photographs,⁸⁵ a criterion which was recently extended to all kinds of works through the interpretation of the Court of Justice of the European Union (CJEU).⁸⁶ A work is original if it is the 'author's own intellectual creation'.⁸⁷ To be eligible for copyright protection, collections of data, tables and compilations must therefore show a sufficient degree of originality in their selection and arrangement.⁸⁸ If the selection and arrangement of the contents of a scientific database are dictated by technical factors or imperatives of accuracy and exhaustiveness, the author can exercise little to no creativity or originality in the choice, sequence and combination of the data in the collection. Scientific databases would therefore not likely meet the threshold for copyright protection, but compilations of scientific articles could be protected. Originality is a question of fact to be determined on a case-by-case basis.

However collections of (scientific) data may also be protectable subject matter under the European *sui generis* database right. Through Article 7 of the Database Directive, as implemented in the legislation of Member States, the maker of a database showing a substantial investment (assessed qualitatively and/or

⁸¹ Kei Iida, Sayuri Imako, Yasutaka Iwamoto, Ong Poh Chuan, Hirohito Katsunuma, Kei Konishi, Junko Kobayashi, Yasuhiko Takada, Takashi Nakazaki, Question Q216B Exceptions to Copyright protection and the permitted Uses of Copyright works in the hi-tech and digital sectors AIPPI National Group: Japanese Group, p. 9.

⁸² L. Guibault, 'Licensing Research Data Under Open Access Conditions under European Law' in D. Beldiman (ed.), *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*, Cheltenham, Edward Elgar, 2013, pp. 63-92.

⁸³ Directive 2009/24/EC of the European Parliament and of the Council of 23 April 2009 on the legal protection of computer programs (Codified version) (Text with EEA relevance) OJ L 111, 5.5.2009, p. 16-22, art. 1(3).

⁸⁴ Directive 96/9 of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20-28, art. 3(1).

⁸⁵ Directive 2006/116/EC of the European Parliament and of the Council of 12 December 2006 on the term of protection of copyright and certain related rights (codified version), OJ L 372, 27.12.2006, p. 12-18, art. 6.

⁸⁶ M. van Eechoud, Along the Road to Uniformity - Diverse Readings of the Court of Justice Judgments on Copyright Work, *JIPITEC: Journal of Intellectual Property, Information Technology and E-Commerce Law*, 2012-1, p. 60-80.

⁸⁷ *Infopaq International A/S v Danske Dagblades Forening*, Case C-5/08, Judgment of the Court, 16 July 2009; *Bezpečnostní softwarová asociace v. Ministerstvo kultury*, C-393/09, Judgment of the Court (Third Chamber) of 22 December 2010; *Eva Maria Painer v. Standard Verlag GmbH*, C-145/10, Judgment of the Court (Third Chamber), 1 December 2011; *Football Dataco v. Yahoo UK Ltd.*, C-604/10 Judgment of the Court (Third Chamber), 1st March 2012.

⁸⁸ T.-E. Synodinou, The Foundations of the Concept of Work in European Copyright Law, in: T.-E. Synodinou (ed.), *Codification of European Copyright Law – Challenges and Perspectives*, The Hague, Kluwer Law International, 2012, pp. 93-113, p. 101.

quantitatively) in either the 'obtaining, verification or presentation of its contents' has the exclusive right to prevent the extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database. Like copyright protection, the *sui generis* database right arises automatically, without any formal requirement, the moment the database is completed or disclosed to the public. The CJEU has given a narrow interpretation of the Directive's requirement of 'substantial investment'. In the landmark cases *British Horseracing Board*⁸⁹ and *Football Fixtures*,⁹⁰ the Court ruled that the term 'obtaining' excludes the costs incurred in the creation of new data (such as generating fixtures lists) from being considered relevant to satisfy the requirement of the substantial investment. Although the costs incurred for creating data are excluded from the calculation of a substantial investment, the costs necessary for the verification of the accuracy of the data and for the presentation of such data to third party users do count in the assessment of whether the investment was substantial.⁹¹ The application of the CJEU principles is particularly complex regarding the distinction between obtaining and creating data and regarding the concrete determination of the investment necessary to trigger the protection. This remains an evaluation that must be made on a case-by-case basis.

Applying the criteria developed by the CJEU to scientific databases, it is unclear whether the majority of research databases meet the formal requirements for the *sui generis* right.⁹² Many collections of data may arguably remain outside the scope of protection because the materials constituting the database are merely created – and not obtained from already existing sources – and the threshold of substantial investment is not reached by further investing either in the verification or the presentation of such content. On the other hand, the investment made by a publisher in the collection, verification and presentation of scientific articles and data sets (Sweet and Maxwell, Taylor & Francis, Reed Elsevier, Beck Verlag and others) would most probably meet the requirement of substantiality, giving rise to protection under the database right regime.

Where the 'obtaining, verification or presentation' of research datasets is deemed a substantial investment sufficient to qualify for protection, the *sui generis* protection confers two transferable rights on the maker of a database: the right of extraction and the right of re-utilisation of substantial parts of the database, which are respectively defined as follows: '(a) 'extraction' shall mean the permanent or temporary transfer of all or a substantial part of the contents of a database to another medium by any means or in any form; (b) 're-utilization' shall mean any form of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or other forms of transmission'. These two concepts have received a broad interpretation from the

⁸⁹ *British Horseracing Board Ltd v William Hill Organization Ltd* (BHB decision), C-203/02, [2004] ECR I-10415

⁹⁰ *Fixtures Marketing Ltd v Svenska AB* (Svenska), C-338/02, [2004] ECR I-10497; *Fixtures Marketing Ltd v Organismos Prognostikon Agonon Podosfairou EG* (OPAP), C-444/02, [2004] ECR I-105449; *Fixtures Marketing Ltd v Oy Veikkaus Ab* (Oy Veikkaus), C-46/02, [2004] ECR I-10365.

⁹¹ See Annemarie Beunen, *Protection for databases – The European Database Directive and its effects in the Netherlands, France and the United Kingdom*, Nijmegen, Wolf Legal Publishers, 2007, p. 137.

⁹² See Mark J. Davison and P. Bernt Hugenholtz, *Football fixtures, horseraces and spinoffs: the ECJ domesticates the database right*, EIPR, 2005-3, p. 113-118, p. 115.

CJEU.⁹³ Recently, the Court of Justice reaffirmed its broad interpretation of the concept of 're-utilisation' in a case involving the display of information generated as a result of a search in a dedicated meta search engine.⁹⁴ The technique employed by a dedicated meta search engine to crawl the targeted databases for specific information, although not identical, is probably comparable to some of the techniques used to text and data mine databases for research purposes: both types of searches make it possible to search the entire contents of that database even if only part of the database is actually consulted and displayed.

Finally, it is worth pointing out that, according to Article 11 of the Database Directive, only natural persons who are nationals of a Member State or who have their habitual residence in the territory of the EU can benefit from the database right. Furthermore, companies and firms are also entitled to such protection if they are formed according to the law of a Member State and have their registered office, central administration or principal place of business within the EU. Article 11.2 clarifies that in case a company or a firm has a registered office only in the territory of the EU, its operations must be substantially and durably linked with the economy of a Member State. In other words, the protection of the *sui generis* database right is not only unique to Europe in that it is conferred only on EU nationals, whether natural or legal persons, but also because there is no real comparable regime of protection for non-original databases outside the EU.⁹⁵

4.3 TDM and the current research exception

Whether and to what extent the use of compilations or databases for purposes of TDM is covered by any relevant exception on copyright or the database right is uncertain. The Database Directive contains a separate set of exceptions for copyright and the database right. With respect to copyright, Article 6(1) contains a mandatory exception on copyright stating that the lawful user of a database may perform, without prior authorisation, any act covered by Article 5 necessary for the purposes of access to the content of the databases and normal use of the content. Article 6(2) allows Member States to provide for limitations on the copyright owner's exclusive rights, including the right to make reproduction of a non-electronic database for private purposes and to use it for the sole purpose of illustration for teaching or scientific research, as long as the source is indicated and to the extent justified by the non-commercial purpose to be achieved.⁹⁶ Since Article 6(2) is optional, Member States have either implemented it in various ways or not at all.⁹⁷

With respect to the *sui generis* database right, Article 8(1) states that 'the maker of a database which is made available to the public in whatever manner may not prevent a lawful user of the database from extracting and/or re-utilising

⁹³ Directmedia Publishing GmbH v Albert-Ludwigs Universität Freiburg, C-304/07, [2009] 1 C.M.L.R. 7.; Apis – Hristovich EOOD v Lakorda AD, C-545/07 [2009] ECRI-1627.

⁹⁴ Innweb B.V. v. Wegener ICT Media B.V. and Wegener Mediaventions B.V., C-202/12, Decision of the Court of Justice, 19 December 2013.

⁹⁵ Among the countries outside the European Union that recognize some protection on non-original databases are South-Korea, Japan.

⁹⁶ L. Guibault and A. Wiebe (eds.), *Safe to be open - Study on the protection of research data and recommendations for access and usage*, Göttingen University Press, Göttingen, 2013, p. 33-34.

⁹⁷ See: Nauta Dutilh, *The implementation and application of Directive 96/9/EC on the legal protection of databases*, Brussels, 2001, Contract ETD/2001/B5-3001/E/72, available at: http://ec.europa.eu/internal_market/copyright/prot-databases/index_en.htm

insubstantial parts of its contents, evaluated qualitatively and/or quantitatively, for any purposes whatsoever'. Article 9 recognises the same optional exceptions on the *sui generis* as in Article 6, but limited to the right of extraction. This means that, where implemented, the substantial extraction of the content of a database is allowed for research purposes, but that no act of re-utilisation can be performed. This restriction, in effect, removes any practical value of the research exception on the database right.⁹⁸

The application of Articles 6 and 9 rests on the concept of a lawful user: only a lawful user may benefit from the exceptions of Article 6(1), 8(1) and 9, while the exceptions listed in Article 6(2) extend to anyone. The concept of 'lawful user' is nowhere defined in the Directive. A literal interpretation suggests that once the rights holder makes the database available to a user, s/he is deemed to be a lawful user.⁹⁹ Access may, however, be conditioned by the terms of use or other contractual agreements set by the rights holder. In such a case, contractual agreement would need to be interpreted in a broad manner. The use of freely available online databases (websites in many instances), even in the absence of any specific terms of use, on the basis of an implied authorisation, may also qualify as a lawful use, as long as the database is published by (or with the consent of) the rights holder.¹⁰⁰

The Information Society Directive also contains an exception on copyright that might be applicable in some cases. Article 5(3)(a) of this Directive allows Member States to provide for exceptions in the case of 'use for the sole purpose of illustration for teaching or scientific research, as long as the source, including the author's name, is indicated, unless this turns out to be impossible, and to the extent justified by the non-commercial purpose to be achieved'. This exception is optional; Member States may decide whether to implement it or not. As a result, Member States have different rules and regulations in this context, where some countries recognise no research exception at all (like The Netherlands and Spain). The assessment made by De Wolf and partners is essentially that the research exception is generally vague and unevenly implemented at national level, which may put some researchers at a disadvantage.¹⁰¹ A second study dedicated solely to the issue of TDM should provide more information on the applicability of the research exception and on the impact of the legal framework on TDM activities.

4.4 Making room for TDM activities under IP law

It appears from the previous section that TDM activities may infringe the rights owner's copyright and/or database right, if done without prior authorisation. The fact the research exception in the Database and Information Society Directives has

⁹⁸ De Wolf and partners, p. 365. See also: A. Beunen, Protection for Databases – The European Database Directive and its effects in the Netherlands, France and the United Kingdom, Nijmegen, Wolf Legal Publishers, 2007, p. 219.

⁹⁹ See Recital 34 offers some guidance: 'Whereas, nevertheless, once the rightholder has chosen to make available a copy of the database to a user, whether by an online service or by other means of distribution, that lawful user must be able to access and use the database for the purposes and in the way set out in the agreement with the rightholder, even if such access and use necessitate performance of otherwise restricted acts'.

¹⁰⁰ See M.M.M. van Eechoud et al., Harmonizing European Copyright Law – The Challenges of Better Law Making, Alphen aan den Rijn, Kluwer Law International, 2009, p. 114.

¹⁰¹ De Wolf and partners, p. 403.

not been implemented in all Member States creates uncertainty within the European scientific community. This may bring about negative repercussions concerning the capacity of researchers to engage in TDM activities on a cross-border basis. Be that as it may, should a measure be adopted to permit acts of TDM, it would need to apply to both the copyright and the database regimes. As discussed in greater detail below, allowing TDM activities to take place for research purposes without fear of infringing IP rights could be achieved in several ways either through an adjustment of licensing practices, through a revised normative interpretation of the 'reproduction right' or through the introduction of an exception on copyright and the *sui generis* database right. Should an exception be introduced in the European legal framework, the legislator would also need to consider whether to ensure that such an exception cannot be overridden through the enforcement of restrictive contractual clauses or technological protection measures.

4.5 Licensing solutions

In late 2012 and early 2013 the European Commission set up a specific Working Group to consider the issue of TDM in the framework of the "Licences for Europe" stakeholder dialogue. While no consensus could be reached among participating stakeholders on either the problems to be addressed or the actions to be taken, publishers presented their own practical solutions to facilitate text and data mining of subscription-based scientific content. As discussed in Chapter 2 this proposal was highly contested by other stakeholders who argued that no additional licences should be required to mine material to which access has been provided through a subscription agreement. The hope is partly that, as governments and funding agencies increasingly demand that the results of publicly-funded research be published under open access conditions researchers will be able freely to access and use an increasing number of databases in addition to the licences offered by publishers in connection with their subscription agreements.

However, a system resting solely on licensing agreements would probably be insufficient to allow TDM to take place in all instances where it would be socially desirable. Firstly, because only a portion of the databases that are interesting for TDM research would be offered as part of publishers' subscription agreement and an even smaller portion would be available under a Creative Commons licence. Without a statutory exception permitting TDM to take place, transaction costs would be too high for parties to negotiate a licence. Secondly, without a statutory exception permitting TDM, there might be little incentives to offer licences under reasonable conditions. In both cases, many databases would remain out of reach of researchers. Thirdly, transaction costs would rise if researchers had to reconcile the terms and conditions of non-standard or non-interoperable licences.

During the 'Licences for Europe' discussions the idea was also put forward to establish a system of voluntary collective licensing whereby permission to text and data mine could be obtained through a collective rights management arrangement. Although attractive in theory, collective licensing would only be workable in practice for the sectors where such collective management systems are already in place, e.g. for texts and musical works. No collective licensing mechanism exists anywhere in Europe for the licensing of rights in databases, and only partial mechanisms exist for

the collective licensing of rights in images and audiovisual works. To allow TDM to occur only through collective licensing would limit and/or delay the application of this solution to certain categories of works only, and/or require the introduction of expensive measures to set up collective mechanisms in other areas of the copyright and database industries.

Normative approach to the reproduction right

The reproduction right in copyright law, as the right of extraction under the database regime, has traditionally received a broad interpretation encompassing any direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part of his/her work. After years of expansive interpretation, it seems timely to ask whether this broad interpretation of the reproduction/extraction right should be reconsidered. Instead of a functional approach to the reproduction/extraction right where all acts of reproduction or extraction that are technically possible fall within the scope of the owner's exclusive right, the legislator could take a normative approach and only recognise protection for acts of reproduction or extraction that actually entail an act of 'expressive' exploitation.

Is TDM a form of copyright or database exploitation that should be under the control of the rights owner? Is TDM (in all its forms) an act of reproduction (and eventually of communication to the public) that affects the interests of the rights owner? American copyright scholars have raised doubts about this insisting that:

*The mass digitization of books for text-mining purposes is a form of incidental or "intermediate" copying that enables ultimately non-expressive, non-infringing, and socially beneficial uses without unduly treading on any expressive – i.e., legally cognizable – uses of the works.*¹⁰²

Arguably, if TDM constitutes *non-expressive, non-infringing, and socially beneficial* types of reproduction, then these should not fall within the ambit of the exclusive right. This would be the normative approach to the definition of the right of reproduction/extraction: if an act of reproduction of a work gives rise to no exploitation of that work, then this act of reproduction should not fall under the control of the rights owner. This normative view of the scope of copyright/database right is rather uncommon nowadays, where directives consistently call for the need to provide a 'high level of protection', which is generally equalised with 'broad protection'. Nevertheless this approach was followed at least on one occasion, by the Dutch government when it implemented Article 5.1 of the Information Society Directive into Dutch copyright law: acts of transient and incidental reproduction that are an integral part of a process or enable a lawful use without having an economic value have been carved out of the copyright owner's exclusive right (Article 13 of the Dutch Copyright Act) instead of having been introduced as an exception.

¹⁰² M. Borghi and S. Karapapa, (2011) Non-display uses of digital works: Google Books and beyond. *Queen Mary Journal of Intellectual Property*, 1 (1), pp. 21-52; Jockers, Matthew L. and Sag, Matthew and Schultz, Jason, 'Brief of Digital Humanities and Law Scholars as Amici Curiae in Authors Guild v. Hathitrust' (June 4, 2013). Available at SSRN: <http://ssrn.com/abstract=2274832> or <http://dx.doi.org/10.2139/ssrn.2274832>; J.H. Reichman and R.L. Okediji, 'When Copyright Law and Science Collide: Empowering Digitally Integrated Research Methods on a Global Scale', 96 *Minnesota Law Review* (2012), pp. 1362-1480; M. Sag, 'Copyright and Copy-Reliant Technology', 103 *Northwestern University Law Review* (2009), 1607-1682.

A shift towards a normative view of the reproduction right could be achieved through an interpretation instrument issued by the European legislator, presumably via a directive. This could be accompanied by a reassessment of the Database Directive, as already done by the European Commission itself in its evaluation report of 2005 of the Directive.¹⁰³ Instead of conferring an exclusive right on the makers of databases, the latter could enjoy a remedy under competition law to stop acts of misappropriation of data by competitors. This would allow acts of extraction and re-utilisation of the content of a database to take place without restriction, if carried out for research purposes.

Exception on copyright and database rights

If the scope of exclusive rights cannot be adapted to reflect a normative view of the right of reproduction/extraction, one option to permit TDM activities could be to introduce an exception on the copyright and database right. As discussed in greater detail below, an exception to copyright and the database right could take either one of two forms: an exception permitting TDM for the purpose of research or an open norm. The two measures have their respective advantages and disadvantages: with an exception on copyright and database right the assessment of whether an act of TDM is lawful would be made *ex ante* by the legislator, while with an open norm the assessment of the lawfulness of an act of TDM would be made *ex post* by the judge. The first option would bring more legal certainty for all parties involved, while the second would bring more flexibility in a fast changing technological environment. Either option must ensure a proper balance between the interests of the rights owner and those of users. In accordance with the international obligations of the European Union under Article 10 of the WIPO Copyright Treaty,¹⁰⁴ the new exception would also need to comply with the requirements of the so-called 'three-step-test', e.g. that the exception be applicable only in certain special cases that do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author.

4.6 Statutory exception

Devising an exception on copyright and database rights allowing for TDM demands the consideration of many factors to ensure that any such exception is indeed not so broad as to unreasonably encroach upon the interests of the rights holders, but not so narrow as to not meet the objective for which it is introduced. The general goal of such an exception would be to encourage the creation of derivative works and transformative uses. Among the elements to consider when defining a new exception for TDM are the subject matter and beneficiaries covered, the scope of the permitted uses, and other conditions of application, such as the payment of compensation. The UK and Ireland are so far the only Member States where the issue of TDM has explicitly drawn the attention of law and policy makers.¹⁰⁵

¹⁰³ European Commission, DG Internal Market and Services Working Paper – First evaluation of Directive 96/9/EC on the legal protection of databases, Brussels, 12 December 2005, available at: http://ec.europa.eu/internal_market/copyright/docs/databases/evaluation_report_en.pdf

¹⁰⁴ WIPO Copyright Treaty (WCT), signed at the WIPO Diplomatic Conference, Geneva, 20 December 1996.

¹⁰⁵ Hargreaves Review, 2011, p. 48;

It is important to point out that the De Wolf study suggests making a distinction in the activities of research that use protected content, between the use as **subject matter for research** and as **tools for research**, which could lead to different conditions of application.¹⁰⁶ According to the authors, using works as subject matter for research would include reproducing works to analyse them or to use them as illustrations, sharing works with colleagues or using 'digital mining techniques to process huge amounts of texts or data'. Under the second type of use, e.g. using works as tools for research, would fall acts like making copies of papers and sharing them with colleagues, extracting data from datasets for analysis and research and organising repositories of scientific works and making these available to the community.¹⁰⁷ According to this, TDM would fall under the first category, e.g. using works as subject matter for research. It is not entirely clear, however, how both categories of use differ from each other in practice and where the boundary lies between the subject matter of research and the tool for research. Is it in the quantity of works gathered into one database or in the technique used to mine? How would this distinction play out within the framework of the database right? What would be the impact of the introduction of a double exception regime on the research community?

To be effective, a TDM exception should not discriminate between types of subject matter covered, between the sources of works or kinds of databases, or between categories of beneficiaries.¹⁰⁸ This approach would coincide with the research exception recognised in Article 5(3)a) of the Information Society Directive and in Article 6(2) of the Database Directive, neither of which discriminate between categories of works, sources or users. Although the Database Directive makes no such restriction (see above), the application of a TDM exception could be limited to works or databases for which the user is already a lawful user, to avoid conferring on the user a right of access to works or databases where none exists.¹⁰⁹

To safeguard the rights owner's interests the scope of the permitted TDM activities could be confined to acts for research purposes. As De Wolf and partners note, the European copyright *acquis* nowhere defines what 'research' is.¹¹⁰ Referring to the definition put forward by the OECD, research and experimental development could be understood as 'creative work undertaken on a systematic basis in order to increase the stock of knowledge, including knowledge of man, culture and society, and the use of this stock of knowledge to devise new applications'.¹¹¹ The burden would lie on the shoulders of the user to prove that the TDM activity was carried out for research purposes.

As noted in the previous chapter, it is debatable whether a TDM exception for research purposes should be restricted to non-commercial activities or whether it should extend to all types of research purposes, including those carried out for

¹⁰⁶ De Wolf and partners, p. 394.

¹⁰⁷ Ibid., p. 355.

¹⁰⁸ In compliance with Directive 2013/37/EU on the re-use of public sector information (OJEU L 175/1 of 27.06.2013), the re-use of data contained in databases maintained by public sector institutions should not cause problems.

¹⁰⁹ The Report of the Copyright Review Committee, Dublin 2013, p. 84.

¹¹⁰ De Wolf and partners, p. 362.

¹¹¹ Frascati Manual 2002, Proposed Standard Practice for Surveys on Research and Experimental Development, OECD, 2002.

profit. Rights owners argue that they should be entitled to reap (some of) the benefits of the added value put on their databases and to a share of financial returns deriving from queries in their databases. A counter-argument holds that confining the exception to non-commercial research activities only may slow down the pace of innovation, for it is not only non-commercial research that generates socially and economically valuable outcomes. Moreover, making the distinction between what is commercial and what is non-commercial may be very difficult in practice, especially in the case of public/private partnerships (PPP), the commercial character of which is often very difficult to ascertain. In any case, a requirement of non-commercial use would follow the lines already set by the Database and the Information Society Directives. Recital 42 of the latter Directive specifies that “when applying the exception or limitation for non-commercial educational and scientific research purposes, including distance learning, the non-commercial nature of the activity in question should be determined by that activity as such. The organisational structure and the means of funding of the establishment concerned are not the decisive factors in this respect”.

Should a TDM exception for research purposes provide for the payment of fair compensation to the rights holder, modelled on the private copying or reprography levy? This would transform the exception into a non-voluntary or statutory licence, where the rights holder may not prevent the use of his work in exchange for the payment of a fair compensation. Such a fair compensation could encourage rights owners to invest in making their databases available in usable, minable formats. On the other hand, calculating what fair compensation is could prove very difficult. Recital 35 of the Information Society Directive explains that the level of ‘fair compensation’ can be related to the possible harm to the rights holders resulting from the act in question. In cases where rights holders have already received payment in some other form, for instance as part of a licence fee, no specific or separate payment may be due. Moreover, the collection and distribution of a fair compensation payment would necessarily occur through a collective rights management, with the drawbacks mentioned above.

To be sustainable and avoid future legislative updates, the wording of the provision should be neutral enough to withstand the passage of time and the likely changes in the technology. The formulation of the exception should seek to define the essence of the process of content-mining in language cast at a sufficiently high-degree of generality that it is not dependent upon a specific view of technology.

A fair question to ask at this point is whether the research exceptions currently contained in Articles 5(3)a) of the Information Society Directive and Articles 6(2)b) and 9(b) of the Database Directive would meet the needs of the European research community by sanctioning TDM activities for non-commercial research purposes. This option would be conditional on at least two important factors: that the provisions be made mandatory on all Member States and that they be unambiguously declared to cover acts of TDM.

4.7 Open Norm

Instead of enacting yet another exception in a closed list of exceptions to deal with the specific issue of TDM, another option could be to introduce an open norm in the copyright and database rights systems. An open norm could introduce flexibility so as to allow TDM activities to take place, along with other types of activities that would pass the test. An open norm could be introduced in copyright and database rights by interpreting the 'three-step test' in copyright law in a balanced way along the lines of the 'Declaration on a Balanced Interpretation of the "Three-Step Test" in Copyright Law'.¹¹² Instead of a restrictive reading of the test that would require exceptions and limitations to be interpreted narrowly, the Declaration suggests 'an appropriately balanced interpretation of the three-step test under which existing exceptions and limitations within domestic law are not unduly restricted and the introduction of appropriately balanced exceptions and limitations is not precluded.'¹¹³ The Wittem Group¹¹⁴ proposed in Article 5(5) of the *European Copyright Code* a slightly adapted version of the 'three-step-test' inspired by the Declaration mentioned above, containing a fourth element requiring that the legitimate interests of third parties are considered. This provision would be applicable as an open norm, in cases similar to but not covered by the exceptions listed in Article 5(1) to (4) of the Code.

Relation with technological protection measures and contract law

If the law were amended to introduce a TDM exception or an open norm, should this provision be declared mandatory? The mandatory character of a provision can normally be decomposed into three elements, to: (1) be implemented across all Member States in order to ensure effective harmonisation of the law; (2) not be subject to contractual overrides; and (3) not be subject to lock-up behind technological protection measures.¹¹⁵ The first element of the mandatory character might be thought non-controversial in the European context; it would certainly represent a step in favour of a 'digital single market'

Regarding the second element, it could be argued that if the European legislator has deemed it appropriate to limit the scope of copyright protection to take account of the public interest, private parties should not be able to derogate from the legislator's intent through contract. This sort of measure is not unprecedented. At the European level, the Computer Programmes Directive and the Database Directive both specify that exemptions provided therein may not be circumvented by contractual agreement. The absence of any such rule was considered briefly during the legislative process leading to the adoption of the Directive. In the second reading of the Proposal for a Directive, Amendment 156 was tabled for the

¹¹² See: http://www.ip.mpg.de/files/pdf2/declaration_three_step_test_final_english1.pdf

¹¹³ Declaration (Aims). See also Section 1 of the Declaration. See: Geiger, Christophe and Gervais, Daniel J. and Senftleben, Martin, The Three-Step-Test Revisited: How to Use the Test's Flexibility in National Copyright Law (November 18, 2013). Available at SSRN: <http://ssrn.com/abstract=2356619> or <http://dx.doi.org/10.2139/ssrn.2356619>

¹¹⁴ *European Copyright Code*, <http://www.copyrightcode.eu/> The Drafting Committee consisted of L. Bently, T. Dreier, R. Hilty, P.B. Hugenholtz, A. Quaendvlieg, A. Strowel and D. Visser. J. Bing, R. Clark, F. Gotzen, E. Mackaay, M. Ricolfi, E. Traple, M. Vivant and R. Xalabarder were in the Advisory Board.

¹¹⁵ De Wolf and partners, 2013, p. 402; L. Guibault, Copyright Limitations and Contracts: An analysis of the contractual overridability of limitations on copyright, The Hague, Kluwer Law International, 2002.

introduction of a new Article 5(6) to the effect that “No contractual measures may conflict with the exceptions or limitations incorporated into national law pursuant to Article 5”.¹¹⁶ At the national level Belgium, Ireland and Portugal have adopted a measure to prevent the use of standard form contracts excluding the exercise of limitations on copyright to the detriment of the user. The downside of making a TDM exception non-overridable by contract would be that it could prevent the emergence of a potentially efficient contractual practice between rights holders and users around the use of databases.

Finally, if the circumvention of technological protection measures were to be made possible to exercise a TDM exception, this could easily be achieved by adding this new exception in the list of exceptions mentioned in Article 6(4) of the Information Society Directive which governs the relationship between the application of technological protection measures and the exercise of certain exceptions.¹¹⁷

4.8 Accessing non-protected databases

Many non-protected datasets (defined as the XL category in the previous chapter) can be found online, since the Internet itself has become a major database,¹¹⁸ where a multitude of actors try to harvest data through mining and analytics techniques for business reasons (customer and audience profiling, marketing, e-commerce, brand reputation, sentiment analysis, etc.), but also for research purposes. For instance, by mining its millions of users’ search queries, Google was able to make accurate predictions about flu outbreaks.

Private actors are not subject to any obligation to open up or share their data with third parties. Even in situations where such data does not enjoy any special copyright or database protection, restrictions on the (re-)use may flow from contractual requirements (in terms and conditions) set by the holder of the data or from the application of technological protection measures. In today’s online environment, the legal validity of online standard form contracts leaves little room for doubt.¹¹⁹ These contracts typically attempt to redefine – outside any intellectual property regime – what is protectable subject matter and therefore legally excludable, and what is not. For instance, licensors may attempt through standard form contracts and technological protection measures to appropriate information that is not protectable subject matter and that should normally remain freely available to anyone. These contracts also attempt to set other conditions of use than those typically admitted under the intellectual property regimes, a practice which can frustrate the objectives that the legislator intended to pursue when defining the scope of protection.

¹¹⁶ European Parliament, Committee on Legal Affairs and the Internal Market, 17 January 2001, PE 298.3685-197.

¹¹⁷ M.M.M. van Eechoud et al., *Harmonizing European Copyright Law – The Challenges of Better Law Making*, Alphen aan den Rijn, Kluwer Law International, 2009, p.

¹¹⁸ The amount of web pages indexed by Google were 1 million in 1998, but quickly reached 1 billion in 2000 and have exceeded 1 trillion in 2008. The rise of social networking applications, like Facebook and Twitter, and of mobile phones becoming the sensory gateway to get real-time data on people from different aspects, further amplifies the already huge web volume. It can be foreseen that Internet of things (IoT) applications will raise the scale of data to an unprecedented level.

¹¹⁹ See : N. Helberger, L. Guibault, M.B.M. Loos, C. Mak, L. Pessers & B. van der Sloot) *Digital Consumers and the Law: Towards a Cohesive European Framework*, Kluwer Law International: Alphen aan den Rijn 2013.

With online user data becoming an important competitive tool for online media platforms and service providers, players try to shield that data by blocking access to it for interoperability, scraping or mining purposes. Reported conflicts mainly relate to access restrictions imposed on potential rivals (as illustrated in the recent conflict between PeopleBrowsr and Twitter about access to the latter's 'firehose'¹²⁰, which resulted in a court order in the United States).¹²¹ Researchers, however, are also confronted with similar practices. A number of reports delivered in the context of the EU's FP7 research programme, for instance, describe difficulties in relation to compliance with terms and conditions (T&C's) set by social network providers for app developers.¹²² Apparently, each platform has specific particularities, which complicates the design and the implementation of new applications or research tools (for instance, for policy simulation in virtual worlds) that rely on different social media spaces. Another complicating factor is the frequent change in T&C's, without any notification, which requires constant re-evaluation and assessment of technical components and, hence, adds significant overheads to the work. In some instances, such change may even risk rendering the whole project objective futile, for instance, if the T&C's change in a way that would not allow for the specific type of use of data that was intended in the project.¹²³

In other words, even when the owner (or holder) of the data cannot exercise copyright or database rights, contractual restrictions or technical protection measures may render TDM more burdensome or even impossible. Could the refusal of a dominant firm to allow a particular use of public domain information, such as a prohibition to 'text and data mine', be found to amount to a violation of Article 102 Treaty for the Functioning of the European Union (TFEU)? If no substitute product for the work or information owned by such an organisation exists, would this

¹²⁰ Twitter's 'firehose' is the massive stream of real-time data that the company makes available for third-party apps to use.

¹²¹ A. Jeffries, "After suing Twitter, PeopleBrowsr wins data access back in settlement – A startup fights for the firehose", *The Verge*, 25 April 2013; <http://www.theverge.com/2013/4/25/4266692/after-suing-twitter-peoplebrowsr-wins-data-access-back-in-settlement>. S.Y. Wahyuningtyas, I. Graef & P. Valcke, "Assessing access problems in online media platforms", *Telecommunications Policy* 2014 (under review).

¹²² This is, for instance, described in more detail in Kosta, E. et al., +Spaces (Policy Simulation in Virtual Spaces) Project: Deliverable D7.4. Legal evaluation report (September 2012), at p.6-13, available from <http://www.positivespaces.eu/>; Kuczerawy, A. et al., Socios (Exploiting Social Networks for Building the Future Internet of Services) Project: Deliverable D3.5. Legal and ethical analysis (August 2012), at p. 20-27, available from <http://www.sociosproject.eu/>; Kuczerawy, A. et al., Deliverable D5.1.5: Final Legal and Ethical Framework for the Deployment of EXPERIMEDIA Testbeds and Experiments (May 2013), available from <http://www.experimedia.eu/>.

¹²³ Twitter recently announced a pilot project through which it will give a 'handful' (sic) of research institutions access to their public and historical data ("Twitter Data Grants"; <https://blog.twitter.com/2014/introducing-twitter-data-grants>). However, the T&C's set by Twitter may deter researchers from actually submitting a proposal. The data grant is open to individuals at single research groups and it is not possible to use the data grant for a cross-partner consortium. Proposals submitted to Twitter will not be treated as confidential and be used by Twitter any way they see fit. Twitter will own copyright to any derivative work they make from a submitted entry: *"You or the owner of the Content still own the copyright in the Content, but by submitting Content to Twitter, you are granting Twitter an unconditional, irrevocable, non-exclusive, royalty-free, fully paid-up, fully transferable, perpetual and worldwide license to evaluate, use, copy, perform, display, publish, transmit, or create derivative works of the Content, or to authorize third parties to evaluate, use, copy, perform, display, publish, transmit, or create derivative works of the Content in any format and on any platform, either now known or hereinafter invented. Twitter will own any derivative works it (or its authorized third parties) creates from the Content. You hereby waive all copyright, trademark, trade secret, patent and other intellectual property right claims you may have against Twitter for evaluating, using, copying, performing, displaying, publishing, transmitting, or creating derivative works of the Content."*

organisation's practice of prohibiting licensees from 'text and data mining' constitute an abuse of the organisation's dominant position?

To amount to a violation of Article 102 TFEU, three conditions must be met. There must be (a) a dominant position, (b) abuse of that dominant position and (c) a resultant effect on trade between Member States. In determining whether an undertaking is dominant on the market, the Commission will consider the position of the parties and of competitors and customers on the relevant product markets and the possibility of market entry and potential competition in product or geographic terms. Furthermore, the undertaking must be found to *abuse* its dominant position in the market. The abuse need not only be aimed at practices which may cause damage to consumers directly, but also at those which are detrimental to them through their impact on an effective competition structure. The refusal to licence is abusive if it has the effect of leveraging the undertaking's dominant position into a secondary market or of preventing or reducing competition from anyone who might wish to use the product or service, and if such refusal is not objectively justified by some proportionate benefit to the competition structure.¹²⁴

The exercise of intellectual property rights is often seen as an objective justification with the result that restrictions under Article 102 TFEU are imposed only in exceptional circumstances.¹²⁵ When deciding whether to compel an information distributor to license its information, a court would first have to define the market in which the parties compete. Unless the user is able to demonstrate that the distributor occupies a dominant position in that market and that its control over the information prevents the user from effectively competing in the market, no access to the work will be granted. As a result, an action which aims at obtaining a compulsory licence is open only to particular classes of users that actually compete or wish to compete in a downstream market. For instance, such an action would hardly be available to individual end-users since they do not 'compete' with the information distributor in the sense of the continental European rules on competition. For the same reason, an action based on the rules of competition law would hardly be available for researchers. A court would also have to enquire about the 'indispensable character' of the work or information held by the dominant undertaking, about the impossibility to duplicate the data or the ideas contained in that work, and about the absence of any other alternative.

In numerous respects, the general criteria of examination developed under the continental European rules on competition are insufficient to address the growing concern about the monopolisation of information. For data produced, collected or paid for by public bodies (so-called public sector information or government data), the EU has already introduced a number of initiatives to support 'open data' and ensure that data like geographical information, statistics, weather data, data from publicly-funded research projects and digitised books from libraries, are available for

¹²⁴ C. Stothers, *Refusal 'To Supply as Abuse of a Dominant Position: Essential Facilities in the European Union'*, [2001] 22 E.C.L.R., 256-262.

¹²⁵ Joint cases C-241/91 and C-242/91, *RTE and ITP v. EC Commission*, 6 April 1995, [1995] 4 C.M.L.R. 718; Case T-504/93, *Tiercé Ladbroke SA v. Commission*, 17 June 1997, [1997] 5 C.M.L.R. 309; Case 7/97, *Oscar Bronner GmbH and Mediaprint Zeitungs- und Zeitschriftenverlag GmbH*, 26 November 1998, [1999] 4 C.M.L.R. 112; Case C-481/01 P(R), *NDC Health Corporation and NDC Health GmbH & Co. KG*, 11 April 2002, [2002] 5 C.M.L.R. 1; Case T-184/01 R II, *IMS Health Inc. v. EC Commission*, 26 October 2001, [2002] 4 C.M.L.R. 2; Case T-184/01 R I, *IMS Health Inc. v. EC Commission*, 10 August 2001, [2002] 4 C.M.L.R. 1.

use and re-use. These initiatives include both legislative measures (such as Directive 2003/98/EC on the re-use of public sector information, revised in 2013, which is built around transparency and fair competition) and non-legislative measures (like the setup of open data portals).¹²⁶ Also, as outlined in Chapter 2, the EU's Open Access strategy is aimed at facilitating use and re-use, in this case of publications and data resulting from scientific research experiments funded at least partially from public funds.¹²⁷

Should such an approach be extended to data held by private entities? Some authors call for a more general regime of (mandatory) openness and interoperability (with open standards) in online environments, to prevent major data holders (one might think of Facebook, Twitter, Google or other online players) "from erecting a fence around its piece of the information commons".¹²⁸ Others suggest that, instead of scrutinising the intent of the monopolist and the harm to the market, the courts should enquire about the motivations that run contrary to the policies behind intellectual property law.¹²⁹ In other words, the courts should not only sanction those situations in which the right owners' anti-competitive behaviour actually harms the market, but also those situations where rights owners enforce their monopolies only or mainly to discourage or prevent others from creating their own works.

4.9 Privacy issues

Discussions on privacy issues and the role of data mining, profiling and data warehousing date back to the 1990s. However, as an ever larger amount of data is being digitized, shared across organisational boundaries and re-used for secondary purposes, privacy and data protection have become even more pressing policy issues.¹³⁰ The proliferation of ubiquitous computing ('Internet of Things', ambient intelligence...) in combination with the growing possibilities for the linking and analysis of data creates the additional challenge that even data which would, taken alone, not raise privacy concerns, may expose wide-ranging impressions of the person concerned, including very sensitive personal data.¹³¹ Sets of correlated data

¹²⁶ For more information, please consult the EC's relevant webpages: <https://ec.europa.eu/digital-agenda/en/open-data-0>.

¹²⁷ See: <http://ec.europa.eu/digital-agenda/en/open-access-scientific-knowledge-0>.

¹²⁸ I. Brown and C.T. Marsden, *Regulating Code: Good Governance and Better Regulation in the Information Age*, MIT Press, 2013; I. Brown and C.T. Marsden, "Regulating Code: Towards Prosumer Law?" (February 25, 2013). Available at SSRN: <http://ssrn.com/abstract=2224263> or <http://dx.doi.org/10.2139/ssrn.2224263>.

¹²⁹ N. Elkin-Koren, 'A Public-Regarding Approach to Contracting over Copyrights', in R. Cooper Drefuss, D. Leenheer Zimmerman and H. First, *Expanding the Boundaries of Intellectual Property*, Oxford, Oxford University Press, 2000, pp. 191-221, p. 215; R.S. Vermet, 'A Synthesis of the Intellectual Property and Antitrust Laws: A Look at Refusals to Licence Computer Software', *Columbia-VLA J.L. & Arts* 1997/22, pp. 27-59, p. 43; and I. Govaere, *The Use and Abuse of Intellectual Property Rights in E.C. Law*, London, Sweet & Maxwell, 1996, p. 149.

¹³⁰ McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity, at p.107; http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

¹³¹ J. Cas, 'Ubiquitous Computing, Privacy and Data Protection: Options and Limitations to Reconcile the Unprecedented Contradictions', in S. Gutwirth et al. (eds.), *Computers, Privacy and Data Protection: an Element of Choice*, Springer, 2011, p.152.

that could be considered insignificant or even trivial can provide intimate knowledge about, for example, life style or health risk, where TDM is applied.¹³²

Current EU rules on data protection provide a high level of cross-sectoral protection for the privacy of individuals, imposing strict limits on the collection and use of personal data. Directive 95/46/EC of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data applies in general to the processing of personal data within the EU. The only exceptions concern public security, defence, State security and the activities of the State in areas of criminal law, and the processing by a natural person in the course of a purely personal or household activity.¹³³ The EU data protection regime will be further strengthened if the draft Regulation – published by the European Commission in January 2012 and currently under debate with the Council and the European Parliament – is adopted later this year.¹³⁴

The collection and processing of personal data for scientific research purposes is also subject to the safeguards imposed by the EU rules, such as the necessity of having a legitimate ground to process such data, the obligation to collect data only as far as it is necessary in order to achieve the specified and legitimate purpose (principle of finality/purpose limitation); the prohibition against collecting more data – and to keep them for a longer period – than is necessary for the purposes for which they are collected and/or further processed (the ‘data minimisation’ principle). Directive 95/46/EC provides only for a limited number of exceptions to these rules and principles for scientific research purposes. Article 13 (2), for instance, allows Member States to restrict the data subject’s right of access when data are processed solely for purposes of scientific research, in cases where there would be no risk of breaching the privacy of the data subject. Generally speaking, researchers who in the context of their projects wish to process personal data have to comply with the rules on data protection. This requirement applies very broadly, to include any information relating to an identified or *identifiable* natural person, whereby it suffices that data can with reasonable efforts be retransformed into personal data.¹³⁵ Even where personal data is made public (e.g. on social media) by the data subject (even manifestly) researchers are not exempt from the requirement of having a legitimate ground for processing such data, which – in most cases – will require the consent of the data subject.

¹³² M. Hildebrandt, ‘Profiling and the identity of the European citizen.’ in M. Hildebrandt and S. Gutwirth (eds.), *Profiling the European Citizen: Cross-Disciplinary Perspectives*, Dordrecht: Springer, 2008, p.304. The aggregation and analysis of digital clinical data from medical records, for instance, may reveal information that help payors and regulators to improve clinical decision making, but may also hold risks for patient privacy.

¹³³ Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31–50.

¹³⁴ Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 25.1.2012, COM(2012) 11 final, 2012/0011 (COD). The articles mentioned in the text refer to the Commission’s proposal, as no major changes were suggested in relation to the aspects discussed in our text by the European Parliament’s LIBE Committee report tabled for plenary, 1st reading/single reading: Report on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (COM(2012)0011 – C7-0025/2012 – 2012/0011(COD)), 21.11.2013, A7-0402/2013.

¹³⁵ Article 29 Data Protection Working Party. Opinion 4/2007 on the Concept of Personal Data (2007).

European research project consortia involved in the mining of information on social networking sites have highlighted the difficulties experienced in seeking the consent of the data subjects, which they consider as very limiting and actually not allowing them to fulfil their original plans (i.e. to use the “abundance of virtual space users”), as they need to ask consent from each and every user.¹³⁶ The requirement for obtaining user consent (and the administrative burden surrounding it)¹³⁷ as well as difficulties relating to the allocation of responsibilities and the principal prohibition of the processing of certain categories of ‘sensitive’ personal data, may hinder the conduct of research and the development of innovative and competing tools involving user data.¹³⁸ The establishment of a general exception for data processing undertaken for scientific or research purposes has been suggested as a potential solution, though it is recognised that this may make it easier for non-scientific researchers to access this type of data.¹³⁹

The Draft Data Protection Regulation partly accommodates those concerns by declaring the processing of personal data (including sensitive data) which is necessary for the purposes of historical, statistical or scientific research as lawful, subject to certain safeguards (Articles 6, 9 and 83).¹⁴⁰ Recital 129 clarifies that scientific research should be understood to include “fundamental research, applied research, and privately funded research”. The general principles that apply to any processing of personal data – such as the ‘collection limitation’ principle, the ‘purpose specification principle’ and the ‘use limitation principle’ – still have to be respected (Article 5). It has been argued that these principles are at odds with the very concept of data mining itself.¹⁴¹ Researchers (or other entities) engaging in data mining wish to accumulate as much data as processable, to generate as much information as possible about individual behaviour patterns and preferences (risking contravention of the ‘data minimisation’ principle). The contents of, and the context in which, this knowledge is going to be applied remains necessarily unclear at the time of collecting the data (potentially falling foul of the ‘purpose specification principle’).

Advanced data analysis technologies, such as TDM, have added a dimension to these ongoing discussions about privacy. The pervasiveness of data collection can

¹³⁶ See, for instance, +Spaces (Policy Simulation in Virtual Spaces) project, Deliverable 7.4 “Legal Evaluation Report”, 2012, p.21; <http://www.positivespaces.eu/>; deliverable available from http://ec.europa.eu/information_society/apps/projects/logos/6/248726/080/deliverables/001_SpacesD74V1_0.pdf.

¹³⁷ Such as filing notifications to the relevant Data Protection Authority/ies, signing of agreements between partners on data protection issues, preparation of consent forms, preparation of privacy notices etc.

¹³⁸ Report of the +Spaces Workshop on the Privacy and Data Protection Framework, Brussels, 8 December 2010 (not published).

¹³⁹ +Spaces (Policy Simulation in Virtual Spaces) project, Deliverable 7.4 “Legal Evaluation Report”, 2012, p.21; <http://www.positivespaces.eu/>; deliverable available from: http://ec.europa.eu/information_society/apps/projects/logos/6/248726/080/deliverables/001_SpacesD74V1_0.pdf. In any case, any exception covering the processing for research or scientific purposes would only be relevant for the duration of the research project and would not be enough to justify the processing of data that may continue for the products of the project after it is over.

¹⁴⁰ Please note that for medical data the Draft Regulation foresees specific rules in Article 81; clinical trials are also subject to the rules adopted by Directive 2001/20/EC of the European Parliament and of the Council of 4 April 2001 on the approximation of the laws, regulations and administrative provisions of the Member States relating to the implementation of good clinical practice in the conduct of clinical trials on medicinal products for human use, OJ L 121, 1.5.2001, p. 34–44.

¹⁴¹ J. Cas, ‘Ubiquitous Computing, Privacy and Data Protection: Options and Limitations to Reconcile the Unprecedented Contradictions’, in S. Gutwirth et al. (eds.), *Computers, Privacy and Data Protection: an Element of Choice*, Springer, 2011, p.141.

easily blur the distinction between sensitive and non-sensitive data, leading to potentially highly sensitive gathering of personal information about individuals.¹⁴² Even in the case of pseudonymous data capture, increasingly powerful and efficient tools for the linking and analysis of large amounts of data allow the re-personalisation of pseudonymous data.¹⁴³

In response to these developments, it has been argued that a fundamental reform of current data protection legislation is needed, requiring a reconceptualization of privacy in terms of access to knowledge instead of data, along with protection against unfair use of that knowledge. Regulatory attention in that case would shift to the use, particularly to the prevention of abuse of personal data or the knowledge gained from them, rather than the technical activities of collecting and processing of data.¹⁴⁴ Moves in this direction might be helpful in avoiding the unintended consequence of measures to protect privacy turning into measures which create further difficulties in the deployment of TDM in scientific research and so further problems for the development of Europe's digital economy.

¹⁴² Ibid., p.146.

¹⁴³ Ibid., p.158.

¹⁴⁴ See, for instance, M. Hildebrandt, Ibid., p.305; J. Cas, Ibid., p.164; V. Mayer-Schönberger & K. Cukier (2013). *Big Data: A Revolution That Will Transform How We Live, Work and Think*, New York-Boston: Eamon Dolan/Houghton Mifflin Harcourt. Also in the US, scholars suggest a legal and regulatory regime that supports privacy through provable accountability to usage rules rather than merely data access restrictions, see: D. Weitzner, H. Abelson, T. Berners-Lee, et al., 'Transparent Accountable Data Mining: New Strategies for Privacy Protection', MIT-CSAIL-TR-2006-007, January 27, 2006, available from: <http://dspace.mit.edu/handle/1721.1/30972#files-area>.

5. Conclusions

From the analysis in this paper, we can draw the following analytical conclusions about TDM and the challenge it presents to policymakers in Europe:

- Text and data mining is an important research technique which is certain to become more important as researchers acquire the skills and the technology to address and investigate datasets of increasing size, complexity and diversity in all media: text, numbers, images, audio files and in any other form.
- TDM represents a significant economic opportunity for Europe. Prolific use of TDM would add tens of billions of Euros in value to the EU's aggregate GDP. This would result chiefly from higher productivity among researchers and from the effects ('externalities') of increased levels of research.
- At present, the use of TDM tools by researchers in Europe appears to be lower, and probably significantly lower, than is the case in the United States and some other countries in the Americas and Asia. This reflects, among other factors, disadvantages created by the European legal framework with regard to TDM.
- The European legislator needs to re-consider and reform the EU's legal framework with regard to copyright, database protection and possibly data privacy, in order to support the international competitiveness of Europe's research base.
- There is a serious risk that Europe's relative competitive position as a research location for the exploitation of 'Big Data' will deteriorate further, if steps are not taken to address the issues discussed in this report. The results of this might well include a loss of talent and a loss of investment to more favourable research locations.

These are the general conclusions of this review. In chapter 4 we outline a range of approaches to achieving different gradations of reform. We recognise the political complexity and likely longer term ambition of some of these proposals, so we set out here a short menu of action points, starting with the immediately available and moving to the most ambitious version of reform, which the Expert Group unequivocally commends.

5.1 Licensing

According to some of Europe's largest scientific publishers, the only response needed to unlock the TDM opportunity is to improve licensing procedures, for example along the lines recently proposed by Reed Elsevier and others. These changes, although with built-in limitations, represent a welcome move from the previously negative stance of some publishers towards TDM. In themselves, however, improved licensing terms for mining scientific publications does not meet the needs of digital age researchers, who require legally reliable research access to many types of database, spread across numerous media platforms, disciplines,

organisations and countries. Some open access publishers have taken another direction and, in the case of PLOS, require authors to sign a data availability statement that guarantees that all the data used in a paper will be publicly accessible to anyone at the moment the paper goes live.

In order to make TDM sufficiently available, Europe needs a new legal framework, either in the form of an exception to copyright and database law, specifically to cover the activities of scientific researchers, or a broader change in the law which would address the needs of text and data miners, along with others caught up in the unintended digital consequences of laws governing European copyright and database protection.

5.2 An exception favouring text and data mining

The case for an exception in copyright and database protection law, applying to text and data mining by scientific researchers, has many merits:

- It plays to Europe's comparative strength in the area of university research, supported by massive scientific research investment at the European level through programmes like Horizon 2020, which is worth approximately €80 billion.
- An exception defined to support scientific research builds upon the existing research exception in the Copyright directive, but could be designed to avoid its shortcomings; ie it could be made mandatory in all Member States and not subject to over-ride by contract or technological protection measures.
- An exception focused upon scientific research poses little risk to the supply of new research data because academic researchers are not motivated directly by the financial gain attached to publication; their career motivations are built around citation and reputation.
- A TDM exception fits with the growing trend towards 'Open Access' academic publishing, which is now well established in most European states, having been embraced by the EU, by national governments, national academic communities and by many publishers, some of whom now enjoy a 'researcher pays' model of remuneration rather than the previously dominant 'reader pays' model. As noted above, more than 40% of scientific peer reviewed articles published worldwide between 2004 and 2011 are available online in open access form¹⁴⁵.
- A surge in TDM among Europe's scientific researchers would undoubtedly spill over into other areas of the public and private data analytics, where additional value would be generated by an emergent generation of highly skilled text and data miners.

What, then, are the shortcomings associated with an exception in copyright law for text and data mining by scientific researchers? The first set of problems concerns

¹⁴⁵ http://europa.eu/rapid/press-release_IP-13-786_en.htm

issues of definition: what is 'scientific' research? What is research? Do we seek to draw a distinction between 'commercial' and 'non-commercial' research in an environment where academics frequently work in partnership (or 'co-creation') with private sector businesses and where today's publicly funded post-graduate research programme is tomorrow's spin-out company? Moreover, as we have argued in the economics section of this report, it does not make sense from a strictly economic point of view to distinguish between the commercial and the non-commercial. The welfare effects of more highly productive research do not recognise the distinction.

A TDM exception applying to *all* scientific researchers, commercial and non-commercial, would avoid most of these problems and would represent a huge improvement on the status quo. But it would surely be more efficient to seek to capture in the European laws which govern copyright and database protection the issue which lies at the heart of these difficulties in defining a TDM exception: how to continue to protect rights-holders against illegal copying of the works upon which their livelihoods and business models depend, whilst avoiding a regulatory overspill of copyright and database law into zones never intended by those who drafted the first copyright laws. This requires us to grapple with the distinction between the illegal copying of 'expressive' works, which sits at the heart of copyright, clear enough in the analogue age, and the mechanical, instrumental copying which is basic to the operation of the Internet and to text and data mining, and which results in 'transformed' outputs which do not compete with (or 'rival') the original works or datasets copied by computers.

It may be possible to capture all of these meanings and intentions in an exception aimed specifically at text and data mining for scientific research, but given the laborious and time-consuming nature of copyright reform and the risk that the language in specific exemptions becomes overtaken by changes in technology and other circumstance, it would surely be better to enshrine the principles described here into a reform with broader effect than an exception covering only text and data mining.

5.3 A strategic reform of copyright and data-base law

If we go back to the foundations of copyright law, we find the English Parliament's 1710 Statute of Anne, stating its goal very broadly as 'the encouragement of learning'. Eighty years later, the first US Copyright Act set as its objective: 'the progress of science and the useful arts.'

Copyright lawyers and other experts have been arguing for many years whether it is possible to distinguish in law between the kind of creative or 'expressive' work, which copyright law is clearly intended to protect from illegal and economically damaging copies, and other forms of copying, which are routine, pervasive and mechanised in the digital age. With TDM, such 'copying' or 'reproduction' does not result in a copy which jeopardises the interests of the rights holder; indeed any resulting output is and should be required to qualify as a 'transformed' product.

In the European debate about copyright, as framed in the closing months of 2013 (and of which this expert review is an element) the question was asked whether reform of the 2001 Copyright Directive is required.

Our examination of the very specific field of text and data mining leads us to the clear conclusion that the answer to this question is: Yes. If Europe is not to hobble its digital economy, it must urgently make a distinction in law between expressive works and the mining of those works by scientific researchers for non-expressive and non-rival purposes. This distinction is required because without it, copyright's original inspiration and motivation, to advance learning, science and the useful arts, is otherwise subverted. By the same arguments, the legislator must re-examine Europe's 'sui generis' database protection directive, to ensure that it too does not present an economically damaging obstacle to scientific research.

So, in concluding, we propose three linked action points:

1. We welcome initiatives to make licensing of works for the purpose of text and data mining easier. In the short term, these will add value to the economy and help to build the skills-base and culture necessary for successful 'big data' research in the digital economy. This activity, however, should be seen as a prologue to legal reform, not an end in itself.
2. A specific and mandatory exception to remove text and data mining for scientific purposes from the reach of European copyright and database law should be considered. This should be regarded as a medium-term amelioration, in the event that our third proposal, below, cannot make timely progress.
3. The best approach to reform is to establish a durable distinction in European law between copyright's longstanding and legitimate role in protecting the rights of authors of 'expressive' works and copyright's questionable role in the digital age of presenting a barrier to modern research techniques and so to the pursuit of knowledge. This initiative should be at the heart of a new copyright directive in Europe, following the consultations currently being undertaken by the European Commission. The legal analysis in this report offers more than one route via which a reform of this kind might be pursued; for example by introducing a suitable 'interpretative instrument' into a new Copyright Directive. We also urge the legislator and the European Parliament to ensure that the currently proposed reform of Europe's data protection laws avoids the unintended consequence of creating further impediments to the work of scientific researchers.

We make these recommendations in the interests of the international competitiveness of the European Union's research base.

Bibliography

Acharya, R. and S. Coulombe (2005) "R&D composition and labor productivity growth in 16 OECD countries", working paper, University of Ottawa and Industry Canada

Depoorter, B., and F. Parisi, 2002. Fair use and copyright protection: a price theory explanation. George Mason law school,
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=259298

Depoorter, B., F. Parisi & N. Schulz. 2002. Duality in property: commons and anticommons
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=224844&rec=1&srcabs=259298&alg=1&pos=2

Depoorter, B., F. Parisi & N. Schultz, 2005, Fragmentation in property: towards a general model
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=224844&rec=1&srcabs=259298&alg=1&pos=2

Filippov, S., *Mapping the Use of Text and Data mining in Academic and Research Communities in Europe*. The Lisbon Council, Brussels (forthcoming).

Gordon, W.J., and R.G. Bone (1999). 1610 Copyright
<http://encyclo.findlaw.com/1610book.pdf>

Guellec, D., and B. van Pottelsberghe (2004) "From R&D to Productivity Growth: Do the Institutional Settings and the Source of Funds of R&D Matter?", CEB Working Paper N° 04/010.

Guellec, D. and B. van Pottelsberghe de la Potterie (2000), "The Impact of Public R&D Expenditure on Business R&D", OECD Science, Technology and Industry Working Papers.

Lemley, M.A. and C. Shapiro., 2007, Patent holdup and royalty stacking. Texas law review
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=923468

Nonneman, W and Vanhoudt, P., "A Further Augmentation of the Solow Model and the Empirics of Economic Growth for OECD Countries." Quarterly Economic Journal of Economics, August 1996, 111(3), pp. 943-53

Schmoch, U., C. Michels, P. Neuhäusler and N. Schulze. 2012. *Performance and Structures of the German Science System 2011*. Berlin: Expertenkommission für Forschung und Innovation.

Tsai, H.-H. 2012. 'Global data mining: An empirical study of current trends, future forecasts and technology diffusions'. *Expert Systems with Applications* 39; 8172-8181.

Tsai, H.-H. 2013. 'Knowledge management vs. data mining: Research trend, forecast and citation approach'. *Expert Systems with Applications* 40; 3160-3173.

An exploration of Google Scholar data

Search results on Google Scholar also provide an indication of growth in TDM. Google Scholar is a widely used Internet search engine for academic publications. Google Scholar employs web crawlers to search the Internet and record information on publications of all types that are either published academically or featured in academic publications. Where possible, it covers full texts. The data presented in this section is based on (manual) data mining of this website.¹⁴⁶ There is one consistent and clear result: the amount of articles referring to “data mining” and “text mining” has been growing rapidly in a roughly exponential growth pattern.

At the outset, some problems in the data need to be acknowledged. First, no detailed documentation of the exact data collection and reporting methods of Google Scholar has been available for this exercise. It is not featured on the Google Scholar website, and there simply was no time to request such information from Google. One inconsistency in the data collected from Google Scholar is apparent: the total score reported for search terms without restrictions on the publication date was often lower than the sum of annual scores of individual years between 1988 and February 2014. This could be because the software restricts the number of very voluminous search results. In any case, this inconsistency is one reason to consider the evidence presented here as preliminary. Incidentally this problem also documents how important it is for data mining for research purposes that comprehensive documentation of the underlying methods is provided along with the data itself.

The first step in this exploration was to enter search terms related to text and data mining, using inverted commas for compound expressions so that only the exact sequence of letters were featured in the search results. The aggregate results for “data mining” was 1.14 million separate items on Google Scholar. “Text mining” brought up 90,400 publications.¹⁴⁷ See Table 1 for an overview of search terms. Results for a number of rough synonyms or overlapping concepts were recorded, to reduce the risk of missing any substantial amount of relevant publications due to varying terminology. Furthermore, Table 1 features search results for terms that are very frequently used in research articles. Results on these general reference terms are useful to develop a sense of the total volume of publications covered on Google Scholar and the share of TDM-related publications in overall research output.

¹⁴⁶ All data was collected from www.scholar.google.nl between 17 February 2014, 20:00hrs and 18 February 2014, 02:00hrs from the same work station / IP address and without protective measures against cookies and personalization of search results.

¹⁴⁷ We restrict ourselves to English language publications throughout.

TABLE 1: The number of search results on Google Scholar for TDM-related terms

Categories	Search terms	Aggregate number without temporal restrictions ^(a)	Sum of annual scores, 1988 to 17 February 2014 ^(b)	Ratio between annual score and aggregate number
Data mining	Data mining	1,140,000	656,888	0.58
	Knowledge discovery ^(c)	378,000	231,474	0.61
	Big data ^(d)	32,100	36,368	1.13
	Knowledge extraction	23,900	--	--
	Information discovery	15,900	--	--
	Data archaeology ^(e)	1,150	--	--
	Information harvesting	1,120	--	--
	Machine learning	1,530,000 ^(f)	--	--
	Analytics	511,000 ^(f)	--	--
Text mining	Text mining	90,400	67,442	0.75
	Text analytics	3,460	--	--
	Content analysis	1,310,000	375,960	0.29
Reference terms	Data analysis	2,310,000	7,926,400	3.43
	Abstract	7,740,000	24,102,200	3.11
	Introduction	6,050,000	27,368,600	4.52
	Survey	4,890,000	20,656,500	4.22
	Empirical	3,150,000	10,763,100	3.42

^(a) This column reports the overall number of search results indicated on Google Scholar if the search term is used without specifying any time frame for the publication date (or any other search restriction).

^(b) This column reports the sum of the number search results for each year of publication between 1988 and 2014 (up to 17 February), which were separately recorded for selected terms.

^(c) Many top hits for "knowledge discovery" also featured "data mining", often even in the publication title. We cannot exclude that this is partially due to the adaptation of Google search results due to previous searches, since all data was collected under the same IP address and without measures to inhibit cookies.

^(d) The top hits for "big data" are mostly commentary rather than applications.

^(e) "Data archeology" resulted in 392 search results.

^(f) Most top hits for this term were unrelated to "data mining" as defined in this report.

Data mining

On Google Scholar, "data mining" features much more frequently than "text mining". Regarding rough synonyms or overlapping concepts for data mining, "knowledge discovery" and "big data" had many additional search results. Other terms closely related to "data mining" either feature less often or bring up many search results that fall outside of the definition of "data mining" or TDM used in this report.

Figure 1 presents annual data on the number of publications on Google Scholar containing “data mining” and important similar concepts. For “data mining”, there is a clear upward trend until 2008 and a downward trend after 2010. As will be shown below, this downward trend in recent years is apparently due to Google Scholar covering fewer recent articles. The proportion of “data mining”-publications to publications containing generally used reference terms increased very consistently. Search results for “knowledge discovery” consistently expand per year up to 2012. “Big data” grew very rapidly since 2011.¹⁴⁸

It is essential to get a sense of the share of TDM-related research publications in all research output. Google Scholar does not feature information on the total number of publications covered, and ‘empty’ searches are not possible. To develop a reasonable reference, we recorded the number of search results that are very frequently used in research publications. See Table 1 and the category ‘reference terms’ for a list of the terms used. Clearly, none of these terms is perfect in the sense that it would be featured and reported on in all relevant publications on Google Scholar. Jointly, results on these terms should provide a reasonable indication of the overall trend in the number of publications featured on Google Scholar.

To identify changes in the share of TDM-related publications in the entire research output, the annual number of search results for “data mining” were divided by the respective results of each reference term. The result was multiplied by 100 to avoid dealing with small fractional numbers.¹⁴⁹ This produces an index that would take a score of 100 if there are as many data mining publications as those for a reference term, 50 if there are half as many data mining publications, and 10 if there ten times as many publications featuring the reference term than data mining. This index is easy to interpret as a percentage figure, even though this is somewhat imprecise as we did not control for the extent of overlap between the search results for different terms.

Figure 2 present the annual index scores. The proportion of “data mining” to the number of results for each of the reference term increased consistently and very rapidly. This holds in particular for those reference terms that typical for empirical research (“data analysis”, “survey” and “empirical”). The apparent decline in the number of publications on “data mining” after 2010 – see Figure 1 – is probably not due to less research activity in this area. It rather seems to reflect a systematic bias of the Google Scholar-database, which features fewer articles for recent years. Google Scholar itself is based on data mining, and it takes time for articles to appear online and for crawlers to gather and incorporate information into the database. In terms of its share in research output, research on data mining is consistently becoming more important.

¹⁴⁸ At least among the top 20 hits for this term, the majority of results on Google Scholar are discussions of the phenomenon rather than empirical applications of the data collection and related analysis methods.

¹⁴⁹ The index is calculated by the equation: $\frac{\text{Score "data mining"}}{\text{Score "reference term"}} \times 100$. All proportions between the number of Google Scholar results for various search terms have been calculated in this manner.

FIGURE 1: Number of search results on Google Scholar for terms related to “data mining”

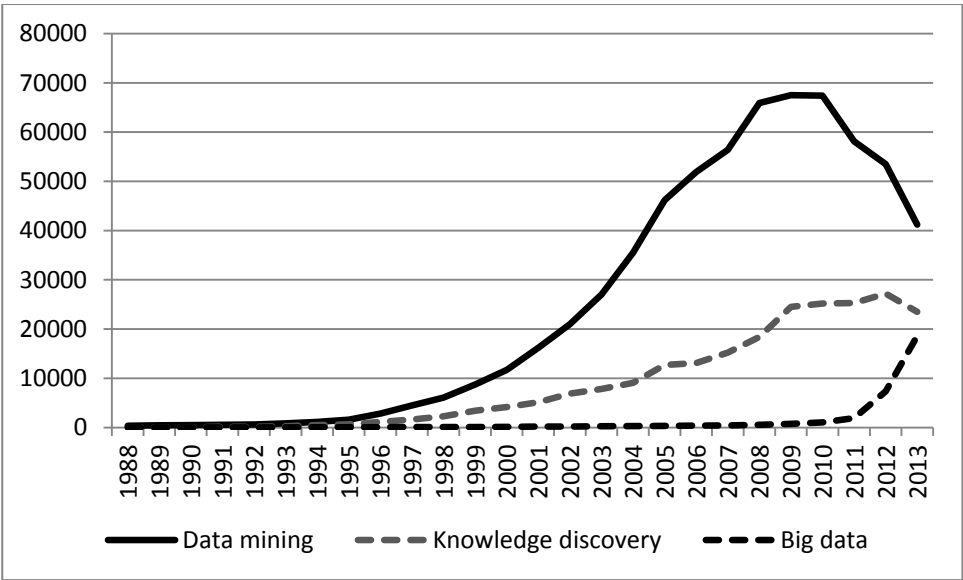
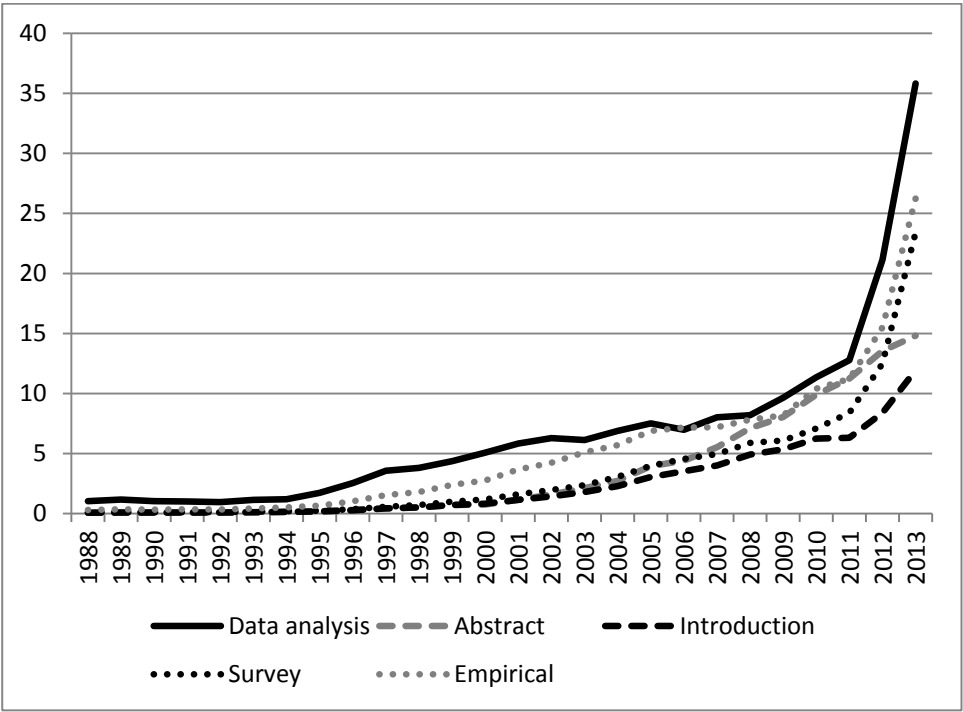


FIGURE 2: Proportion of search results for “data mining” and reference terms



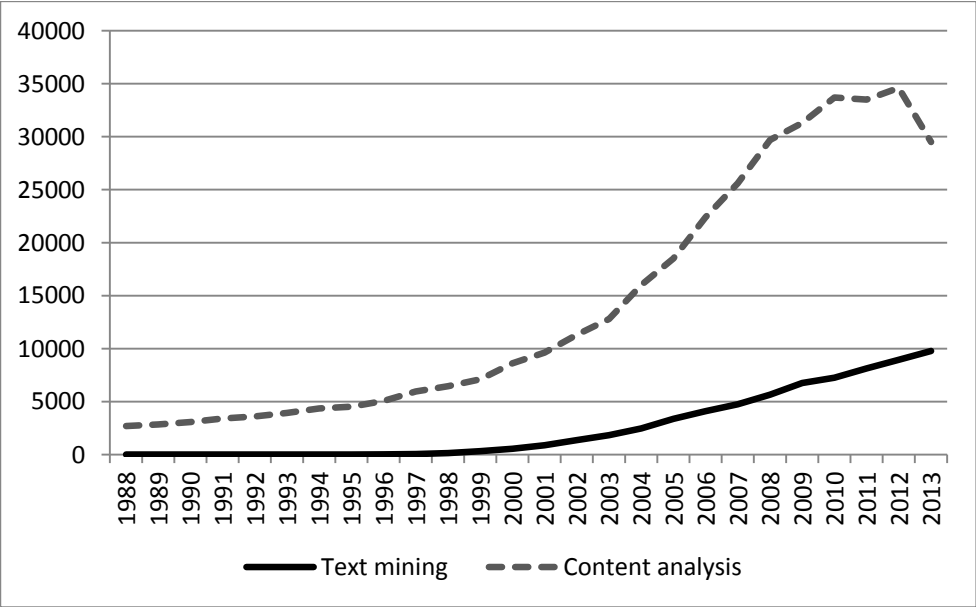
Notes: For each year, the figure shows the number of search results for “data mining” on Google Scholar divided by the number of search results for frequently used terms in research articles multiplied by 100, that is: $\frac{\text{Score "data mining"}}{\text{Score "reference term"}} \times 100$.

Text mining

For “text mining”, the only apparent rough synonym is “text analytics”. This overlapping concept produced few search results and is not addressed in detail here. Text mining is a subordinate concept to “content analysis”, the quantitative analysis of qualitative (textual) information. Figure 3 reports the absolute counts of search results for “text mining” and “content analysis”. Figure 4 presents the index value of the proportion. The indication is that text mining has become much more important within this category of research over the last two decades.

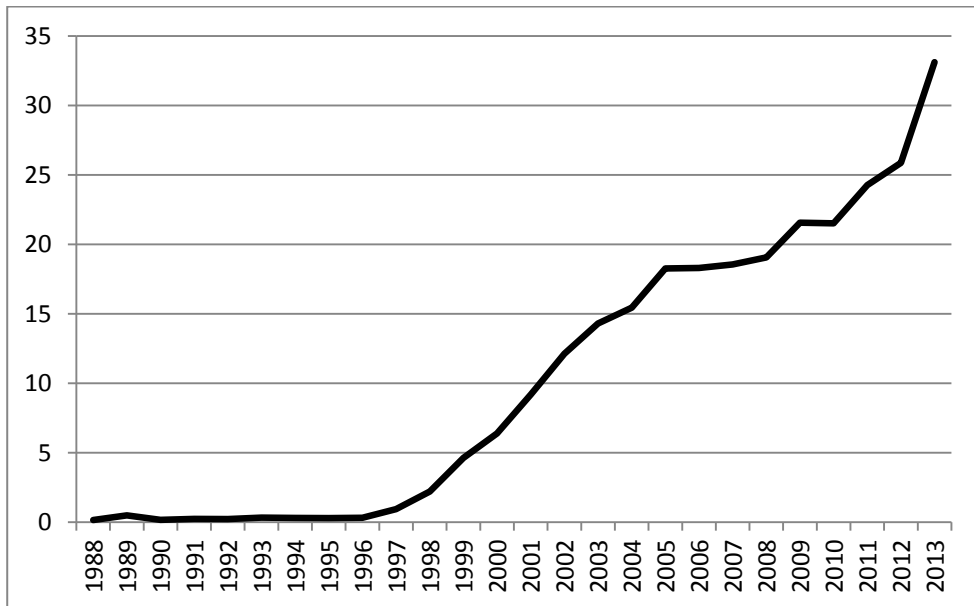
Regarding the proportion of “text mining” in research output at large, there is an even more rapid growth pattern than for data mining – see Figure 5. Another way to show this is by estimating the proportion of research articles that feature “text mining” and “data mining” – see Figure 6. The relative frequency with which text mining featured is up from not much more than 1 in 200 for 1996 to almost 1 in 4 for 2013.¹⁵⁰

FIGURE 3: Number of search results for “text mining” and “content analysis”



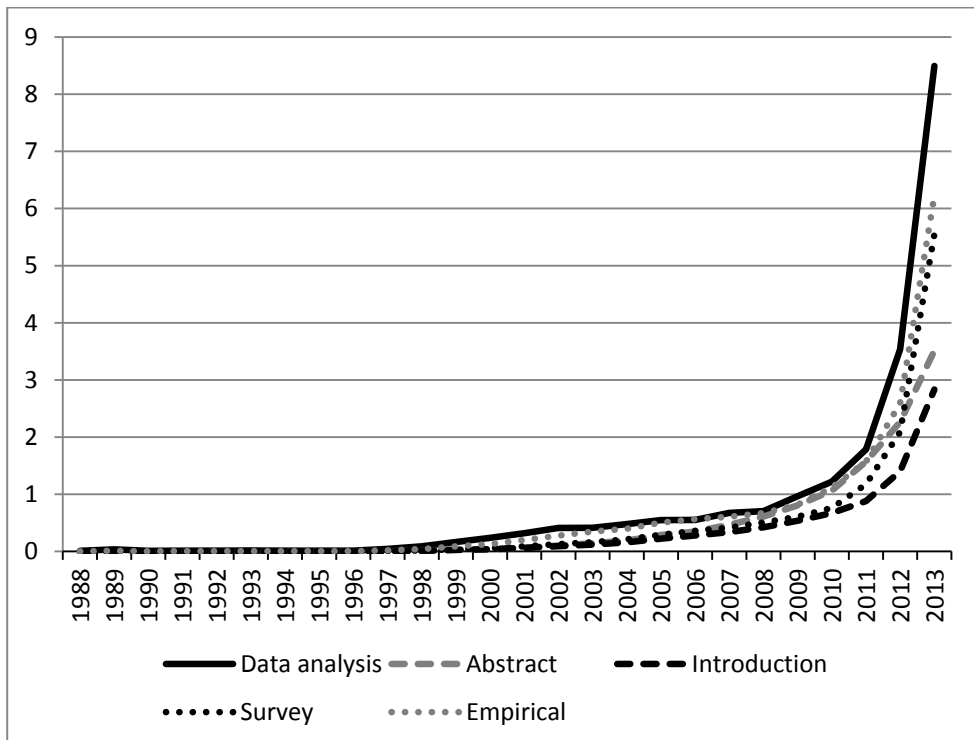
¹⁵⁰ The popularity of the expression “text and data mining” does not influence this proportion greatly, since it is used relatively infrequently. It produces 1,190 results without temporal restrictions and 162 results for 2013.

FIGURE 4: Proportion of search results for “text mining” and “content analysis”



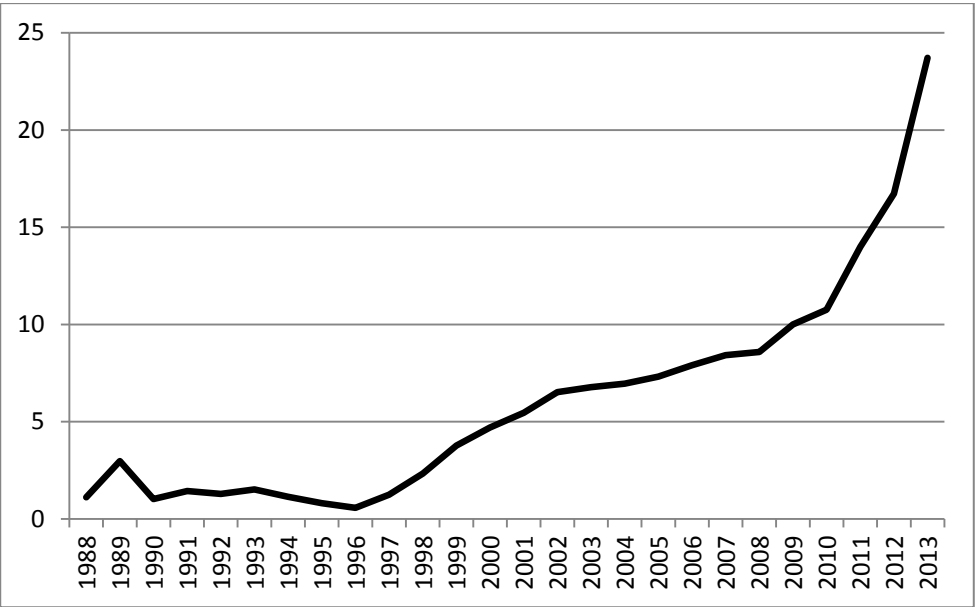
Notes: For each year, the figure shows the number of search results for “text mining” on Google Scholar divided by the number of search results for “content analysis” multiplied by 100, that is: $\frac{\text{Score "text mining"}}{\text{Score "content analysis"}} \times 100$.

FIGURE 5: Proportion between search results for “text mining” and reference terms



Notes: For each year, the figure shows the number of search results for “text mining” on Google Scholar divided by the number of search results for frequently used terms in research articles multiplied by 100, that is: $\frac{\text{Score "text mining"}}{\text{Score "reference term"}} \times 100$.

FIGURE 6: Proportion between search results for “data mining” and “text mining”



Notes: For each year, the figure shows the number of search results for “text mining” on Google Scholar divided by the number of search results for “data mining” multiplied by 100, that is: $\frac{\text{Score "text mining"}}{\text{Score "data mining"}} \times 100$.

Summary of the analysis of Google Scholar data

This basic exploration of search results on the search engine Google Scholar demonstrates that TDM accounts for an increasingly large share in total research output. Growth rates over recent years have been high. This outcome is consistent with the secondary data from Thomson Reuter’s Web of Science discussed earlier. Data mining related research already makes up a surprisingly large share of publications covered on Google Scholar. Text mining is less frequently referred to in academic work but growing even more rapidly.

On a more general level, this use of Google Scholar data demonstrates the logic of derivative and transformative use of digital data. Google Scholar itself is based on data mining, and we mined that data within the technical infrastructure developed by Google. Last but not least, the credibility of the data used here and research opportunities would be greater if some additional services were available, such as a sufficiently detailed documentation of the underlying methods.

European Commission

**Standardisation in the area of innovation and technological development,
notably in the field of Text and Data Mining
- Report from the Expert Group**

Luxembourg: Publications Office of the European Union

2014 — 75 pp. — 17.6 x 25.0 cm

ISBN 978-92-79-36743-4

doi 10.2777/71122

How to obtain EU publications

Free publications:

- one copy:
via EU Bookshop (<http://bookshop.europa.eu>);
- more than one copy or posters/maps:
from the European Union's representations (http://ec.europa.eu/represent_en.htm);
from the delegations in non-EU countries (http://eeas.europa.eu/delegations/index_en.htm);
by contacting the Europe Direct service (http://europa.eu/europedirect/index_en.htm) or
calling 00 800 6 7 8 9 10 11 (free phone number from anywhere in the EU) (*).

(*) The information given is free, as are most calls (though some operators, phone boxes or hotels may charge you).

Priced publications:

- via EU Bookshop (<http://bookshop.europa.eu>);

Priced subscriptions:

- via one of the sales agents of the Publications Office of the European Union
(http://publications.europa.eu/others/agents/index_en.htm).

Text and data mining (TDM) is an important technique for analysing and extracting new insights and knowledge from the exponentially increasing store of digital data ('Big Data'). TDM is useful to researchers of all kinds, from historians to medical experts, and its methods are relevant to organisations throughout the public and private sectors. TDM represents a significant economic opportunity for Europe. Prolific use of TDM would add tens of billions of Euros in value to the EU's aggregate GDP. At present, the use of TDM tools by researchers in Europe appears to be lower than in its main competitors. There is a serious risk that Europe's relative competitive position as a research location for the exploitation of digital data will deteriorate further, if steps are not taken to address the issues discussed in this report.

Studies and reports

978-92-79-36743-4

