the Lisbon council
think tank for the 21st century

Text and Data Mining for Research and Innovation

Interactive policy brief

Issue 20/2016

What Europe Must Do Next

By Sergey Filippov and Paul Hofheinz

Sergey Filippov is associate director and Paul Hofheinz is president and co-founder of the Lisbon Council, a Brussels-based think tank. This paper follows on Mapping Text and Data Mining in Academic and Research Communities in Europe, originally launched at the OECD Expert Workshop for Knowledge-Based Capital in 2014.

In the age of the Internet, the volume of published scientific articles continues to grow rapidly, driven by intensified awareness of the value of research to the extension of knowledge and its application through innovation.[1] Estimates are hard to come by, but some calculations show that there are as many as 60 million academic articles in circulation today, with roughly 1.5 million new ones added each year, many of them in an increasingly wide range of disciplines, which scholars are increasingly expected to have knowledge of as well.[2] Google Scholar, a search engine that indexes the full metadata of literature across an array of formats and disciplines, links to around 165 million documents according to some calculations, or roughly two billion pages.[3] Indeed, the volume is so large that no one – not even the high-powered servers in massive data farms spread around the world – can put a precise figure on it.

This poses a dilemma for scholars. Most academic literature is created on the principle that new research should build on and extend work that was done before – a practice which often means that the start of any meaningful academic inquiry is an investigation into and analysis of the sum of all preceding findings on the topic at hand.[4] But how can researchers take account of this gigantic maze of extant literature? How can they possibly keep track of more than 1.5 million new articles each year, many of them highly technical, and some of them presented only at conferences in hard-to-reach cities, or available online in sometimes obscure journals?

The opinions expressed in this interactive policy brief are those of the authors alone and do not necessarily reflect the views of the Lisbon Council or any of its associates.

This interactive policy brief seeks to make knowledge more accessible through online circulation and interactive features, such as hotlinks to articles cited in the footnotes and a web-friendly format.

[1] The authors would like to thank Dirk Pilat and Jeremy West of the OECD; Ian Hargreaves, Stéphanie Lepczynski and Elizabeth Miles of the Lisbon Council; and members of the Roundtable on Demystifying Text and Data Mining in a Copyright Context in the European Parliament, hosted by MEPs Therese Comodini Cachia, Miapetra Kumpula-Natri and Catherine Stihler. As always, any errors of fact, judgment or omission are the authors' sole responsibility.

[2] Arif Jinha, "Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence," Learned Publishing, 23(3): 258–263, 2010. The 60 million figure is based on extrapolating existing trends forward from the latest available year, 2010, to the present.

[3] Enrique Orduna-Malea, Juan M. Ayllón, Alberto Martín-Martín, and Emilio Delgado López-Cózar, "Methods for Estimating the Size of Google Scholar," Scientometrics, 104(3): 931-949, 2015.

[4] The process is known in the academic field as "literature review," and is often presented as the opening chapter of any new paper.

'New technologies make analysis of large volumes of text and other media potentially routine.'

Fortunately, there are new techniques to which skilful researchers can turn – most notably text and data mining, an advanced algorithm-based reading method in which large volumes of text and data from existing articles can be analysed, categorised and sorted to detect patterns and extract meaningful information for a wide variety of purposes. Text and data mining enables users to spot and summarise relationships that weren't visible before and to analyse and uncover patterns and relationships in diverse databases where connections were previously not easy to establish. Put simply, it's a way of clearing out much of the academic underbrush. Scholars can quickly determine which articles are or are not relevant to their research without having to read them. And others can build interesting and meaningful new avenues of insight, finding patterns where others didn't see them, and performing the task that scholars have always performed – taking the available facts and putting them together in ways that yield new insights and advance the overall knowledge of mankind.

Why we need a universal mandatory exception for text and data mining

On 09 December 2015, the European Commission proposed a mandatory exception for research in EU copyright legislation for "public interest research organisations to carry out text and data mining of content they have lawful access to, with full legal certainty, for scientific research purposes."

The initiative was welcomed by many if not most in the research community, though some expressed concern about the language used, not least the term "public interest research organisation," which is disputable and therefore likely to be a target for legal challenge. The second part of the text – "for scientific research purposes" – is also open to misinterpretation, both in the scope and meaning of the word "science" (does that include machine-analysis of images for market research purposes?) and even the term "research" which might be argued to exclude experimental study without any pre-defined objective, or a formulated hypothesis. A proposed exception for "non-commercial" research is problematic, too, given ever closer partnerships and collaborations between publicly funded research institutions and companies – an objective aggressively pursued by public policy in recent years.

Like the *Ligue des Bibliothèques Européennes de Recherche* (LIBER), OpenForum Europe and many others, we believe that the only workable and justifiable solution is the least ambiguous one: a harmonised, mandatory exception at the EU level covering all text-and-data-mining activities, for any purpose, commercial and non-commercial, and an exception that cannot be overriden by a contract and is applicable to all rights holders – corporate, individual, public and private. Like these advocates, we believe that greater adoption of text and data mining is an important prerequisite for informing and energising the European economy in the digital era. Policy should support the embrace of this technique rather than seeking to resist it.

[5] The 2014 paper was prepared for the Organisation for Economic Co-operation and Development (OECD) Expert Workshop for Knowledge-Based Capital, which convened in Paris on 27 May 2014. It later served as input for the Expert Group on Text and Data Mining, convened by the European Commission's directorate-general for research and innovation. Subsequently, the paper generated much interest in the wider research and policymaking community – where it served as early evidence of a fast-growing and increasingly important development within the European research community.

[6] Sergey Filippov, *Mapping Text and Data Mining in Academic and Research Communities in Europe* (Brussels: Lisbon Council, 2014).

[7] The paper has been widely quoted in much recent research on text and data mining. See, *inter alia*, Organisation for Economic Co-operation and Development, *Data-Driven Innovation Big Data for Growth and Well-Being: Big Data for Growth and Well-Being* (Paris: OECD, 2015). Ibid., "Making Open Science a Reality," *OECD Science, Technology and Industry Policy Papers*, No. 25 (Paris: OECD, 2015). Science Europe Working Group on Research Data, *Text and Data Mining and the Need for a Science-Friendly EU Copyright Reform* (Brussels: Science Europe, 2015). Christian Handke, Lucie Guibault and Joan-Josep Vallbé, "Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research," *Social Science Research Network (SSRN)*, 20 May 2015. Éanna Kelly, "Researchers to Take on Publishers over New EU Copyright Laws," *Science|Business*, 07 May 2015.

[8] See, also, European Commission, *Report from the Expert Group on Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining* (Brussels: European Commission, 2014).

> ‘European scholars are forced on occasion to outsource their text-and-data-mining needs to researchers elsewhere.’

Two years ago, the Lisbon Council conducted a large-scale research project on the use of text and data mining in academic and research communities in Europe.[5] We found then that while European academics displayed a healthy interest in the new technique, there were increasing signs of quicker uptake elsewhere in the world.[6] Now, less than two years later, we decided to revisit the topic to see if the situation had changed.[7] The question did not come up haphazardly, either.[8] On 23 March 2016, the European Commission announced a “Public Consultation on the Role of Publishers in the Copyright Value Chain and on the ‘Panorama Exception,’” in which the European Union’s executive arm quietly proposed the extension of so-called “neighbouring rights” to the realm of scientific, technical and medical publications.[9] Many believe this bold proposal flies squarely in the face of other policy initiatives to help European researchers remain at the forefront of global knowledge creation and to make research more accessible within Europe for broader scientific use.

Given text and data mining’s potentially strategic role in global scientific progress, we believe that European policy frameworks should promote and facilitate adoption of this advanced technique (see the box on page 2 for more). It is not enough to strengthen and extend the arrangements of the pre-digital age, taking steps to add ambitiously conceived new property rights to the already well-established rights upon which the analogue economy is based. To the contrary, we must look objectively at research-community requirements in this field. And we must do this particularly if we are to count on the historic role of innovation and innovation-led growth to drive forward a better, more balanced European economy, where leadership at the cutting-edge of knowledge translates easily into better social and economic outcomes for all.

As in our 2014 paper, we employed a multilevel analysis to assess Europe’s position *vis-à-vis* other global regions in the exploration of advanced text-and-data-mining techniques.[10] This was done using quantitative (publication and patent analysis) and qualitative approaches (semi-structured interviews). The results were striking.

Among the key findings:

- Around the world, scholars continue to generate an impressive volume of articles on text and data mining each year, but the weight of activity is shifting towards new markets. Over the last decade, Asia has replaced the European Union as the world’s leading centre for academic research on text and data mining as judged by number of publications.

- In the six-year period covering 2011-2016, Asian scholars’ share of academic publications in the field rose to 32.4% of all global publications, up from 31.1% in 2000. The EU’s global share fell to 28.2%, down from 38.9% in 2000. North America remained in third place at 20.9% due to the relatively small size of the three-country region (though the United States continued to dominate in the separate country rankings with a No. 1 global position).

- The surge of interest from China signalled shifts elsewhere within Asia as well. As recently as 2000, Japan and Taiwan led Asia with 12.6% and 7% of all global text-and-data-mining-based publications, respectively, making up the core of the region’s

[9] The consultation itself consists of 23 multiple-choice questions. We believe this is a fundamentally disingenuous way of going about introducing an entirely new right that will (one way or the other) alter the balance of power between scientific, technical and medical publishers and the broader research community. We believe future copyright revisions should give broader scope for impact assessment. This paper is intended to provide context to the broader conversation that should be taking place around this proposed new right.

[10] Filippov, op. cit.

> ‘The volume of published academic research continues to grow rapidly.’

presence in the field. After a steady rise in interest, China now leads. On its own, it accounted for 11.7% of all global publications in 2015, up from zero in 2000. This gave China a No. 2 finish in the country rankings, second only to the U.S. China’s ranking within Asia is now No. 1.

- China also led the global growth in the number of patents pertaining to data mining. While the number of patents granted by the U.S. Patent and Trademark Office (USPTO) remained relatively stable over the past decade, the number of patents granted for data-mining-related products by the State Intellectual Property Office of the People’s Republic of China (SIPO) rose to 149 in 2015, up from just one in 2005.

- In this context, the patent data for Europe is not relevant or comparable, as software design (including text-and-data-mining applications) is not patentable on the continent. Others caution that high Chinese patent rates could simply be a proxy for heavy prioritisation of patent filings in the 12th and 13th five-year plans. But the fact remains: Chinese researchers and organisations are patenting text-and-data-mining procedures at a faster rate than any other country in the world. It suggests that Chinese researchers attach a growing priority to the potential use of this new technique for stimulating scientific breakthroughs, disseminating technical knowledge and improving productivity throughout the scientific and technical community.

- Some of the fastest growth and greatest interest was seen in relative newcomers: India, Iran and Turkey. Having shown virtually no interest in text and data mining as recently as 2000, the Middle East is now the world’s fourth largest region for research on text and data mining, led by Iran and Turkey.

- Large European scientific, technical and medical publishers have added text-and-data-mining functionality to some dataset licences, but the overall framework in Europe remains slow and full of uncertainty. Many smaller publishers do not yet offer access of this type. And scholars themselves complain that existing licences are too restrictive and do not allow for generating the advanced “big data” insights that come from detecting patterns across multiple datasets stored in different places or held by different owners.

- Legal clarity also matters. Some countries apply the “fair-use” doctrine, which allows “exceptions” to existing copyright law, including for text and data mining. Israel, the Republic of Korea, Singapore, Taiwan and the U.S. are in this group. Others have created a new copyright “exception” for text and data mining – Japan, for instance, which adopted a blanket text-and-data-mining exception in 2009, and more recently the United Kingdom, where text and data mining was declared fully legal for non-commercial research purposes in 2014. Some researchers worry that the UK exception does not go far enough; others report that British researchers are now at an advantage over their continental counterparts.

‘The policy frameworks deployed should promote and facilitate adoption of text and data mining.’

- New technologies make analysis of large volumes of text and other media potentially routine. But this can only happen if researchers have clearly established rights to use the relevant techniques, supported by the necessary skills and experience. Broadly speaking, the European ecosystem for engaging in text and data mining remains highly problematic, with researchers hesitant to perform valuable analysis that may or may not be legal. The end result: Europe is being leapfrogged by rising interest in other regions, notably Asia. European scholars are even forced, on occasion, to outsource their text-and-data-mining needs to researchers elsewhere in the world, as has been reported repeatedly in past European Commission consultations. Anecdotally, we hear stories of university and research bureaux deliberately adding researchers in North America or Asia to consortia because those researchers will be able to do basic text and data mining so much more easily than in the EU.[11]

This interactive policy brief has four parts. In part I, we look at the transformative role of text and data mining in academic research. In part II, we present the findings from an advanced bibliometric analysis, benchmarking the role of European research against global competitors. In part III, we present key findings of a patent analysis, highlighting the global trends (ex. Europe) in the evolution of text and data mining. And in part IV, we reflect on the prospects for an enabling policy in the text-and-data-mining field within the broader European political and economic context.

Table 1. Academic research in the 21st century

	Classic research	Research in the 21st century
Idea for a research paper	Human intelligence, the researcher’s intuition	Human intelligence, supported by technology that is able to automatically identify trending or emerging themes and topics, nuances and gaps
Literature review and hypothesis formulation	Manual search for relevant academic publications; the researcher needs to read them all to understand their relevance to the research project	Technology systematically reviews and extracts relevant scientific data in published literature, classifies it according to multiple dimensions and connects the dots
Data and methodology	Observations and semi-structured interviews; data modelling; statistical analysis of quantitative data, e.g. correlation, causation	Data analysis from many different dimensions and angles, data categorisation. Machine learning: creation of new data models without being explicitly programmed. Discovery: data mining can be used to create databases that can be themselves mined
Analysis of results	Analysis of the past and the present	Analysis of the past and the present provides a basis for predictive analysis of the future

[11] See *Ligue des Bibliothèques Européennes de Recherche/ Association of European Research Libraries (LIBER), Text and Data Mining: The Need for Change in Europe* (The Hague: LIBER, 2014). We have heard similar accounts in background briefings with scholars, though, as the European legal regime for this activity remains uncertain, few are willing to speak on the record about the “work arounds” they are concocting to be able to do research that is free, easy and legal elsewhere.

‘We must look objectively at research-community requirements in this field.’

Text and Data Mining in Academic Research

Text and data mining reads and analyses documents using natural language-processing algorithms. It recognises similar concepts, facts, terms, relationships and assertions.[12] Technology can do preparatory work at every stage of scholarly inquiry – literature review, hypothesis formulation, data analytics – enabling researchers to spend more time and energy on analysis and innovative deduction. The result is dramatically enhanced productivity of the digital-era researcher, who is more likely using text and data mining to produce genuinely new insights. The differences between manual and technologically-enhanced academic research are illustrated in Table 1 on page 5.

But problems arise not simply over the newness of the technique, but the confusion of the legal regime within which research takes place. Here, Europe has long taken a cautious path, whether in balancing the rights of researchers to obtain and use data or the terms upon which rights to use it are made available by publishers. The result is that the right or the incentive to research thoroughly, using rapidly developing digital tools, is more constrained in Europe than in many other parts of the world – and that is even before the introduction of an entirely new “neighbouring right” offering additional rents to rights holders and publishers in the scientific, technical and medical field.[13]

In Europe, it is common for researchers who have the right to read an article not to have the right to subject that article to computer analysis.[14] In order to use text and data mining techniques, a European researcher will often require the express permission of the copyright owner – most often, an academic publisher.[15] This is much less likely to be the case in countries where copyright law allows a general “fair-use” defence against a charge of copyright infringement. Fair-use regimes are found in Israel, Republic of Korea, Singapore and Taiwan – all examples of successful high-tech economies – but most significantly in the U.S., where fair use is enshrined in the 1976 Copyright Act. In most instances, fair use permits limited copying and distribution without permission of the copyright holder or payment for uses such as commentary, search engines, criticism, news reporting, research, teaching, library archiving and scholarship.[16]

To be sure, some academic publishers are developing frameworks to make it easier for researchers to practice text and data mining in Europe. For example, Elsevier, the Netherlands-based global publishing house, has adopted a licence–based approach which both formalises the right to mine into academic agreements and delivers a way in which researchers can gain access to its application-programming interface (API) through a self-service portal. Similarly, Springer, Germany’s global academic publisher, grants text-and-data-mining rights to subscribed content to researchers via their institutions, provided the purpose is non-commercial research. The selection and refinement of desired articles can be conducted by using existing search methods and tools, such as PubMed, Web of Science and Metadata API, among others. Wiley, a U.S.-headquartered global publishing company, grants subscribers and other lawful users the right to text and data mine online content for non-commercial purposes. Users must use an approved API service such as CrossRef’s TDM service or a Wiley API. Normally, the text-and-data-mining clause is included in all new subscription agreements and users

[12] Text-and-data-mining software, often combined with artificial intelligence, can uncover particular patterns and relations in databases where information is uniformly formatted. In contrast, text mining works with unstructured natural language text; it extracts meaningful information and insights that can be used for a wide variety of purposes. Together, text and data mining enables users to analyse data from many different dimensions and angles, to categorise it and to summarise the relationships identified.

[13] As of May 2016, these “neighbouring rights” were still only a possible proposal being discussed. The European Commission conducted an open “public consultation” on this proposed new right, for which this paper was prepared.

[14] “Copyright law originating from the time of the printing press does not prohibit a human from reading and making associations between the facts contained in content,” concludes a Universities UK and UK Higher Education International Unit report. “But it does prohibit a computer programme from undertaking the same work. Due to the fact a computer must make a copy of an entire copyright work in order to perform the same activity, the process of data mining becomes subject to copyright law.” See Universities UK and UK Higher Education International Unit, *European Commission’s Stakeholder Dialogue ‘Licences for Europe’ and Text and Data Mining* (London: Universities UK, 2013).

‘Progress can only happen if European researchers have clearly established rights to use the relevant techniques, supported by the necessary skills and experience.’

are not charged an additional fee for text and data mining, provided the scope remains purely non-commercial.

Although this is a welcome development on the part of academic publishers, its value should not be overstated. It falls well short of providing a norm binding all publishers to a common approach upon which researchers can rely. Even where mining rights are available, they involve complex and varied restrictions, often requiring use of a publisher-controlled portal, where researchers must register as a developer and get an API key.[17]

There are further complications when it comes to the right to include images or data visualisations in mining-based analyses. These are not available in API by default which means that researchers are offered only text mining, not text and data mining. Publication or analysis resulting from text and data mining of the databases of all three aforementioned publishers may include quotations from the original text of up to 200 characters, or 20 words, or one complete sentence. In contrast, permissions to reproduce images must be negotiated on a case-by-case basis.

Elsevier’s text-and-data-mining agreement explicitly forbids efforts to “utilise the TDM output to enhance institutional or subject repositories in a way that would compete with the value of the final peer reviewed article, or have the potential to substitute and/or replicate any other existing Elsevier products, services and/or solutions.”[18] Likewise, Wiley’s terms and conditions prohibit the user from performing “systematic or substantive extracting for the purposes of creating a product or service […] that has the potential to substitute and/or replicate any other existing Wiley product, service and/or solution.”[19]

Academic Interest in Text and Data Mining: A Bibliometric Analysis

In the Lisbon Council’s 2014 paper on text and data mining, bibliometric analysis – statistical analysis of written publications – showed that more and more scholars were producing academic publications on text and data mining across a variety of disciplines, a quantifiable indication of rising interest within the academic community.

In 2016, we revisited the data and found important changes. In 2000, EU researchers produced the most publications on text and data mining, accounting for 38.9% of all such publications that year. Asian scholars ranked No. 2 with 31.1%. Today, Europe and Asia have changed places, with EU researchers responsible for 28.2% of text and data mining publications in the 2011-2016 period and Asia at 32.4%. Scholars from Chinese universities, virtually absent from the field 15 years ago, drive this change.

The data behind these conclusions derived from key word searches in two databases: Google Scholar and the ScienceDirect database of Elsevier. In Google Scholar, we searched for the key words “data mining” between 1990 and 2015, first in the text of

[15] The exception is the United Kingdom. The UK government introduced changes in copyright law that became effective on 01 June 2014. An exception in copyright law now allows for “computational analysis” to be carried out legally on material under copyright. The need for such exception was articulated in the Review of Intellectual Property commissioned by Prime Minister David Cameron and led by Prof. Ian Hargreaves. See: United Kingdom Government, *Digital Opportunity: A Review of Intellectual Property and Growth. An Independent Report by Professor Ian Hargreaves* (London: UK Government, 2011).

[16] More recently, some countries – such as the United Kingdom and Japan – have adopted specific carve outs from IP protection rules for text and data mining, stating explicitly in law that text and data mining of a data set does not constitute a copyright infringement. Jean-Paul Triaille, Jérome de Meeûs d’Argenteuil and Amélie de Francquen, *Study on the Legal Framework of Text and Data Mining* (Namur: De Wolf and Partners, 2014).

[17] Diana Cocoru and Mirko Boehm, *An Analytical Review of Text and Data Mining Practices and Approaches in Europe* (London: Open Forum Europe, 2016).

[18] Elsevier’s Text and Data (TDM) Mining Licence.

[19] Wiley’s Text and Data Mining Agreement.

> ‘The U.S. continues to dominate in the country rankings with a No. 1 global position.’

the publication and secondly in the title of the publication.[20] This was intended to give us a sense of the focus and scope of the articles surveyed. If the term “data mining” was mentioned in passing in the body of the article, we concluded that it indicated general interest in, or at least awareness of, data mining techniques. But when the term was used in the title, we counted it as a direct indication of the content and focus of the publication. Academic publications with “data mining” in the title are usually very focused on this research technique and discuss either new methods of data mining or results of the application of this methodology to particular cases.[21]

We found that while overall interest in text and data mining remains high, the pace of new publications is slowing. This could indicate a saturation of interest and, perhaps, a trend towards specialisation. The evolution of both categories of publication is shown in Table 2 and Chart 1 below.

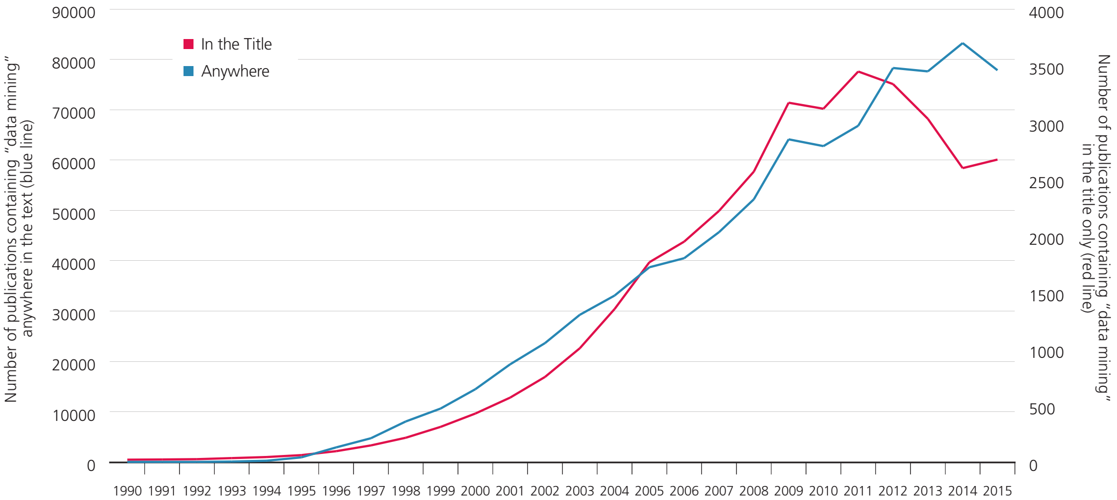
Table 2. Key statistics on text and data mining publications (1990-2015)

Publications	1990-2015	2015	Peak year in 1990-2015
“Data mining” in the title only	37,124	3,460	3,700 (2014)
“Data mining” anywhere in the text (including the title)	786,339	60,100	77,600 (2011)

Source: Lisbon Council calculation, using data from Google Scholar

The blue line denotes the number of all publications containing “data mining” anywhere in the text. The number rose to 60,100 in 2015, up from 450 in 1990, with a peak of 77,600 in 2011. Here we can see a certain saturation of general interest in this research methodology. The red line – denoting the number of publications with “data mining” in the title of the article – is more revealing. The number of these more

Chart 1. Number of publications on data mining (1990-2015)



[20] The methodology deployed relies heavily on the path-breaking article by Hsu-Hao Tsai, a researcher at National Chengchi University in Taiwan, who analysed research trends and forecasts of data mining from 1989 to 2009 by examining 1,181 articles with the heading “data mining” in topic in the Social Sciences Citation Index (SSCI) database of Thomson Reuters. In our 2014 paper, we cited and relied almost exclusively on Mr. Tsai’s results. In this paper, we were able to calculate our own results based on the ScienceDirect database of Elsevier and Google Scholar. Hsu-Hao Tsai, “Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions,” *Expert Systems with Applications*, 39(9): 8172–8181, 2012.

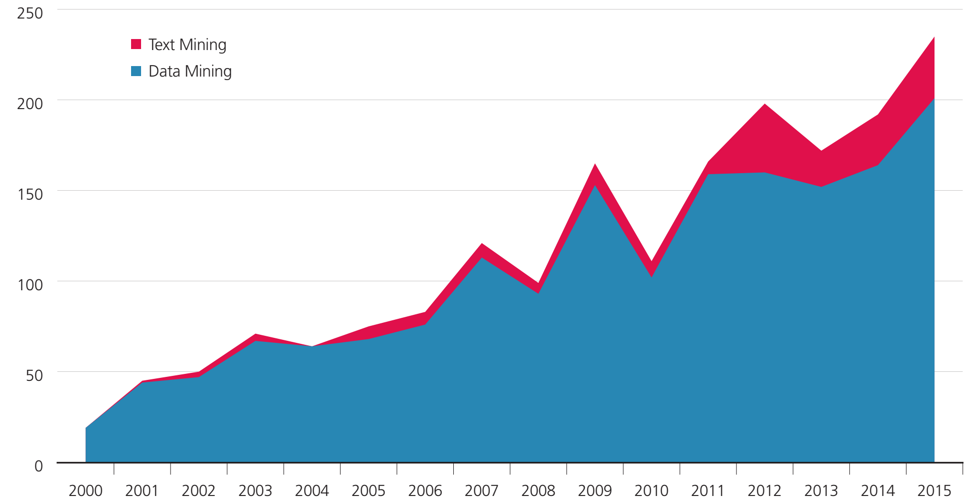
[21] The search was conducted on 27 April 2016. We used the search function “with the exact phrase” – to exclude articles on data in the mining industry. Patents and citations were excluded from the search, too. Note that due to the dynamic nature of web pages indexation, the numbers may slightly differ from day to day.

> ‘Problems arise not simply over the newness of the technique, but the confusion of the legal regime within which research takes place.’

specialised publications rose to 3,460 in 2015, up from just one in 1990, with a peak of 3,700 in 2014. The small decline noted for 2015 may or may not represent a sustained change in the direction of the curve.

A search of Elsevier’s ScienceDirect database of academic publications with the words “data mining” in the title by year of publication yielded similar results.[22] In addition,

Chart 2. Number of publications on text and data mining (2000-2015)



Source: Lisbon Council calculation, using data from Elsevier’s ScienceDirect

Table 3. Distribution of geographic sources of published articles on text and data mining, 2011-2016: Top 10 countries

Rank	Country	Number of publications	Share
1	United States	189	17.8%
2	China	124	11.7%
3	Taiwan	69	6.5%
4	India	63	5.9%
5	**Spain**	**59**	**5.5%**
6	Korea	38	3.5%
7	**United Kingdom**	**36**	**3.4%**
8	Australia	35	3.2%
9	Canada	34	3.2%
10	**Italy**	**32**	**3.0%**

Source: Lisbon Council calculation, using data from Elsevier’s ScienceDirect
Note: The number of publications is rounded off.

[22] ScienceDirect hosts over 12 million pieces of content from 3,500 academic journals and 34,000 e-books in scientific, technical and medical research. These publications are peer-reviewed articles in academic journals, conference proceedings published in academic journals, book chapters and educational materials. Certain publications with “data mining” in the title have the word “text” in the title, too. For more information, visit ScienceDirect

'European scholars complain that existing licences are too restrictive and do not allow for generating advanced "big data" insights.'

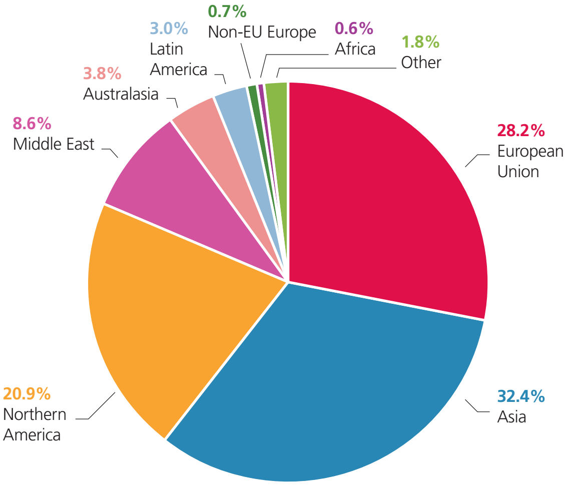
we performed the same exercise for the term "text mining" for publications that didn't appear in the "data mining" search query.[23] Over the 16-year period from January 2000 through December 2015, we found 1,866 publications in total – 1,682 publications on data and text mining and 184 on text mining only. Remarkably, on an annualised basis, we found a 12-fold annual rise in text and data mining publications over the same period – with 235 publications during 2015, up from 19 in 2000.[24] See Chart 2 on page 9 for a graphic representation of this growth.

One of the advantages of the ScienceDirect database is that it provides the academic affiliation of all authors. We used this information to identify the countries whose scholars authored the publications.[25] In total, the number of publications on text and data mining published in ScienceDirect amounted to 1,063 from 01 January 2011 to 01 May 2016. Researchers from 69 countries published or co-published on data mining. Researchers affiliated with US universities topped the list with 188 publications. China occupied the No. 2 slot with 124 publications, and Taiwan came third with 69.

Regarding the country-based rankings, three European countries appear in the top 10 over the 2011-2016 period. They are Spain at No. 5 with 59 publications; United Kingdom at No. 7 with 36 publications and Italy at No. 10 with 31. Table 3 on page 9 lists the country ranking and global shares based on bibliometric analysis.

It is telling that in all non-EU countries in the top 10 list – among them Australia, Canada, China, India, Republic of Korea, Taiwan and the U.S. – data mining without express permission is not explicitly prohibited for academic research, being allowed

Chart 3. Geographic sources of published articles on text and data mining (2011-2016) Regions of the world



Source: Lisbon Council calculation, using data from Elsevier's ScienceDirect
Note: "Other" – affiliation of authors is unknown.

[23] To be sure, publications referring to the mining industry – e.g. "recent data from mining" – were eliminated from analysis. Corrigendum or withdrawn articles were excluded too.

[24] In addition, we performed a search using 2016 as a publication year. As of 01 May 2016, 85 publications in "data (and text) mining" and 15 publications in "text mining" were retrieved, 100 in total. Some of these articles are to be published in forthcoming issues of academic journals, others are accepted as corrected proofs – and all of them are already available electronically.

[25] In exploring the national origin of research publications, we logged the "academic nationality" of the researcher, not their birth nationality. For instance, a publication written by a Chinese scholar who is affiliated full-time with a U.S. university is considered as originating from the U.S. Where an author has multiple affiliations, we identify the country of the principal affiliation. In case of co-authorship, when scholars are affiliated with universities or research institutes located in different countries, we use the rule of thumb widely used in academia and assign a proportional weight to each country. For instance, if four professors based in four different countries write a joint article, each of these countries would receive 0.25 points. The working assumption is that all co-authors contribute equally. This approach is needed to differentiate between sole and co-authorship; or, in our case, between co-authorship by scholars affiliated with more than one university. In fact, the rise of co-authorship is one of the results of the "publish or perish" mindset. See: Andrew Plume, and Daphne van Weijen, "Publish or Perish? The Rise of the Fractional Author…," *Research Trends*, 38, September 2014.

‘China accounted for 11.7% of all global publications on text and data mining in 2015, up from zero in 2000.’

in most cases under a “fair dealing,” fair use or similar exception.[26] In the top three EU countries, the situation is different. In the UK, there is a text-and-data-mining exception as of 2014; and in addition to that, researchers affiliated with UK universities participate actively in international research collaborations, including those with U.S. and Asian universities. Spanish and Italian universities have built a strong skills base in data analytics. But with regards to the copyright regime, their national legislations do not provide for a text- and data-mining exception.

We later aggregated individual country shares to obtain regional shares (see Chart 3 on page 10 for the results). It revealed that the top three global regions are responsible for more than 80% of publications on text and data mining: Asia (32.4%), led by China, Taiwan and India; the EU (28.2%), led by Spain, the UK and Italy, and Northern America (20.9%), consisting mostly of the U.S. and Canada. The EU’s No. 2 rank, with 299 publications, is a sum of 24 national shares (researchers from Croatia, Estonia, Latvia and Malta did not produce any publication on text and data mining). The top six countries – Spain, Italy, United Kingdom, France, Germany and Portugal – accounted for two-thirds of the result (see Table 4 on page 12).

The Middle East is becoming more prominent, led by Iran, Turkey and Egypt. Today, it is the No. 4 region, responsible for 8.6% of all publications worldwide on text and data mining. Region No. 5 is Australasia consisting of Australia and New Zealand, accounting for 3.8%. Latin America, with a 3.0% share, is led by Brazil, Mexico and Chile. The “Non-EU Europe” region is No. 5 (0.7%), led by Serbia, Switzerland and Russia. Africa comes last, with only three countries – Mauritius, South Africa and Swaziland – accounting jointly for 0.6%.

Oddly, we also found that, while there is increasing interest in discussing text and data mining throughout the global academic community, the tools deployed in this area of research indicate methodological caution. The vast majority of articles about text and data mining – including the growing number of topic-related literature reviews – are being carried out using the traditional keyword search provided in publishers’ databases and not with the more advanced techniques of text and data mining to which these articles are ostensibly devoted.[27] This indicates that there is vast scope to improve our knowledge of text and data mining by applying the technique to the study of text and data mining itself.

Patent Analysis

Patent analysis in academic research is a commonly used indicative measure of the innovativeness of countries and industries, allowing the detection of trends surrounding the developments and value of particular technologies.

Text-and-data-mining tools usually rely upon software and the rules governing the patentability of software vary from place to place. So, software is patentable in the U.S. and in many other jurisdictions, but not in Europe where Article 52 of the European Patent Convention states that “programmes for computers” are not inventions for the purpose of granting European patents. For this reason, patents cannot be considered

[26] Australia and India could be possible exceptions, though the law is not entirely clear in either place. Australia and India both rely on a “fair-dealing” exception, which is the norm in most British Commonwealth countries. Under that rule, exceptions to copyright on specific works can be automatically granted if the proposed use falls within a specific category of acts subject to exemption. But the law can be vague, especially when it encounters new and novel ways of using content, such as text and data mining. Because of this, some scholars have concluded that – based on the prevailing fair-dealing rules in place – text and data mining without prior authorisation is “probably” not permitted in Australia or India. See Handke, Guibault and Vallbé, op. cit.

[27] Tsai, op. cit.

‘Research is itself a creative act.’

Table 4. Geographic sources of published articles on text and data mining in the European Union (2011-2016)

Rank	Country	Number of publications	Share	
			In the EU	In the World
1	Spain	59	19.8%	5.56%
2	United Kingdom	36	12.0%	3.39%
3	Italy	32	10.6%	2.98%
4	France	29	9.7%	2.73%
5	Germany	26	8.6%	2.41%
6	Portugal	18	5.9%	1.66%
7	Belgium	16	5.3%	1.48%
8	Greece	14	4.7%	1.32%
9	Poland	12	4.0%	1.14%
10	The Netherlands	12	3.9%	1.10%
11	Czech Republic	9	2.9%	0.82%
12	Sweden	8	2.8%	0.79%
13	Denmark	6	1.9%	0.53%
14	Slovenia	6	1.8%	0.52%
15	Austria	5	1.5%	0.43%
16	Finland	3	1.1%	0.32%
17	Ireland	3	1.0%	0.27%
18	Slovakia	2	0.7%	0.19%
19	Bulgaria	2	0.5%	0.15%
20	Lithuania	1	0.4%	0.13%
21-22	Hungary	1	0.3%	0.09%
21-22	Romania	1	0.3%	0.09%
23	Luxembourg	1	0.2%	0.05%
24	Cyprus	1	0.1%	0.02%
	Total EU	**299**	**100%**	**28.1%**

Source: Lisbon Council calculation, using data from Elsevier’s ScienceDirect
Note: The number of publications is rounded off; due to rounding off, the sum of individual country share is higher than the total of 299.

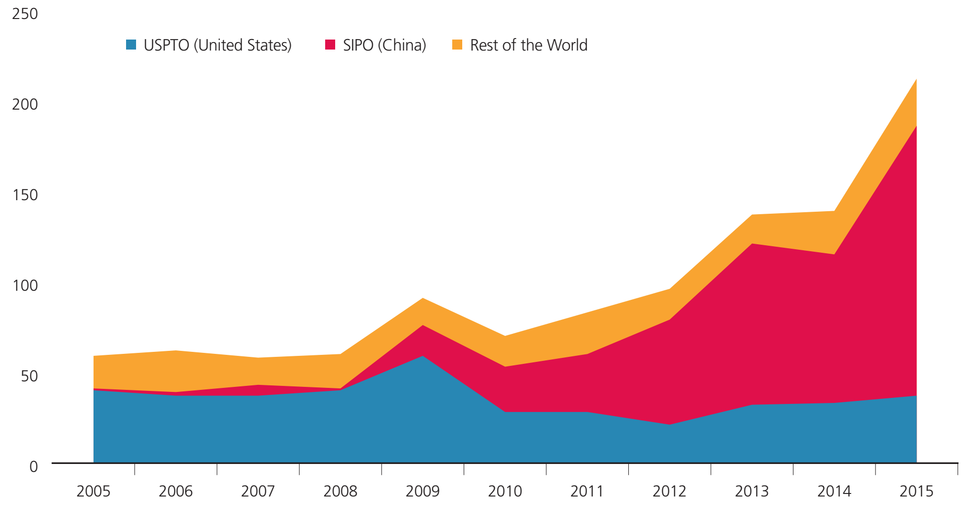
‘Policy should support the embrace of this technique rather than seeking to resist it.’

as a direct proxy for measuring Europe’s performance or level of activity in the field of text and data mining and it cannot be used to make comparisons between countries where patent regimes differ sharply. However, it is still worth examining global trends in patents which refer to text and data mining in order to identify specific trends within other, non-European jurisdictions.

For this task we relied on the EspaceNet patent database of the European Patent Office, using its worldwide database functionality that allows searching for information among more than 90 million published patents from over 90 patent-granting authorities. We counted the number of actual patents awarded and not the number of applications filed. In the 2014 Lisbon Council policy brief on text and data mining, we noted a strong upward trend in the number of patents in text and data mining granted by the State Intellectual Property Office of the Peoples’ Republic of China (SIPO). We found then that SIPO had already overtaken the U.S. Patent and Trademark Office (USPTO) as the leading patent-granting authority in this area. Our updated figures tell us that this trend has continued – and deepened.

Within the 2005-2015 period, we identified 1,067 patents worldwide with the term “data mining” in the title. The number grew to 212 in 2015, up from 59 in 2005 (see Chart 4 below). The number of patents granted by USPTO remained relatively stable throughout the period, ranging from 30 to 40 a year. In contrast, the number of patents granted by SIPO exploded to 149 in 2015, up from just one in 2005.

Chart 4. Patents granted in data mining (2005-2015)



Source: Lisbon Council calculation, using data from European Patent Office’s EspaceNet

> ‘Technology can do preparatory work at every stage of scholarly inquiry: literary review, hypothesis formulation, data analytics.’

Quo Vadis Europe?

As the world moves ever deeper into the digital age, text and data mining has emerged as an important new method for delivering more comprehensive and better evidenced academic research. It is today vital to the process of establishing the current state of knowledge quickly and efficiently in almost any field of inquiry. That is how text and data mining contribute to the waves of innovation which continue to arise from the impact of digital communications technologies. These forces are now widely acknowledged to be fundamental to economic prosperity and well-being of nations in the 21st century.[28]

The research carried out for this paper confirms that European scholars are aware of the importance of text and data mining, but indicators of comparative performance and other data suggest that their approach is hesitant. “There are far more TDM friendly regimes in operation in the U.S., Asia, Canada and the UK,” writes LIBER, the *Ligue des Bibliothèques Européennes de Recherche,* which unites more than 400 leading European research libraries, adding “a number of European-based research projects have already outsourced their content mining to the U.S.” We suggest that the main source of this trend is the complexity of the legal and contractual regimes governing text and data mining in Europe. These regimes differ from country to country and from publisher to publisher. This is the absolute opposite of a unified digital market of the kind imagined in the wider debate around a digital single market for the EU. It is a significant brake on Europe’s ability to return to – and remain at – the forefront of cutting-edge global innovation in the years to come. And it is not a situation that is helped by the creation of an entirely new right, generating additional rents for practices occurring rent free – and against a backdrop of full legal certainty – in all other advanced economies in the world.

What the two Lisbon Council studies on text and data mining, separated by two years, show is that the U.S. retains its long-held leadership position in the study of text and data mining, that Europe is slipping further behind and that Asia is rising strongly. In Asia, relatively recent entrants to the league tables of advanced innovation – China and India – challenge the leadership of the traditional Asian centres of excellence in the field – Taiwan, Korea and Japan. And there is increasing evidence that European research institutions are being forced to reach outside of Europe to build better teams for text-and-data-mining-related consortia – not because the foreigners’ researchers are more able, but because their laws are smarter and more straightforward than the ragged patchwork of rules which apply in Europe.

Most arguments around policy in the area of intellectual property tend to focus on questions of motivation: are creators and inventors incentivised to create by having stronger intellectual property laws? Or do those strong intellectual property laws themselves sometimes hamper innovation?[29] A discussion on this point is beyond the scope of this paper. But one conclusion is reasonably clear: researchers who engage in text and data mining are adamant that what they are producing is original work, the product of their own minds and insight.[30] It may rely on “facts” enshrined in a database or external article, but the value arises from the interpretation researchers are able to put on the facts themselves, sometimes relying on machines to help them see

[28] Organisation for Economic Co-operation and Development, *Data-Driven Innovation: Big Data for Growth and Well-Being* (Paris: OECD, 2015).

[29] Ian Hargreaves and Paul Hofheinz (eds.), *Intellectual Property and Innovation: A Framework for 21st Century Growth and Jobs* (Brussels: The Lisbon Council, 2012).

[30] “TDM is concerned only with the extraction of non-copyrightable objects (facts and data),” writes LIBER in a position paper. “It therefore makes no sense when drafting a law to address the technical shortcomings of the current copyright framework to limit that solution to non-commercial use.” See LIBER, op.cit.

‘The surge of interest from China signaled shifts within Asia as well.’

more deeply. This, in turn, provides a platform for new insights and invention. In other words, research is itself a creative act. If this creativity is choked by clumsy regulation or licensing systems designed to protect established businesses from innovation by others, the wider economic effect is bound to be negative. Unless Europe’s intellectual property laws help us move towards more “knowledge creation,” Europe risks increased dependency upon living off knowledge generated elsewhere.

Within the context of the international debate about intellectual property laws, the doctrine of fair use, deployed in countries like Israel, Republic of Korea, Singapore and the U.S., has enabled researchers and the organisations that employ researchers to see deployment of text and data mining as an acceptable business risk, in legal terms. European researchers and the organisations which employ them, by contrast, face a maze of restrictions, which are in themselves often detrimental, but which in aggregate generate confusion and undermine the self-confidence of the research community.[31] The publishing industry itself continues to put a premium on “licensing,” preferring to retain control – and potentially the right to collect rents – over every use, reuse and even derivative use of any material they may have helped put into the public arena. The result is a minefield of hidden obstacles for European researchers, forcing them to avoid the kind of technology-driven value creation which is routine in North America and surging in Asia.[32]

In recent years, European policymakers have lurched from one side of the issue to the other. They call for reform in one context.[33] And advocate retreat in another.[34] To be clear, this is not so much the same people delivering different messages at different times. Rather, this is the European bureaucracy fighting with itself. On one side is the part of the community that supports and sustains research – universities and libraries – as well as some regulators and activists supporting Europe’s faster move into the digital economy, including think tanks like this one. On the other side is the more political part of the European policymaking scene – politicians who respond to direct pressure from the media and lobbyists – as well as certain parts of the digital economy-based bureaucracy. Their motives may be sincere but their worldview seems dominated by powerful rights holders fighting a series of rear-guard actions against unavoidable innovation and change.

And in the middle of all of this sits the European citizen. Where does his or her interest lie? The evidence tells us this: Europe has a strong interest in remaining at the forefront of the now pervasively digital knowledge-based economy. Only an enabling regime, which sits comfortably with the digital age, where artists are fairly compensated and researchers are fully empowered, will be adequate to the very real challenges of the times. While text and data mining originates from universities, research institutes and commercial data-wranglers, there are no objective reasons to limit this breakthrough methodology to “ivory towers.” This 21st century methodology should step out of universities and find its application for research and innovation across the modern economy and in society at large. This calls for a modern, contemporary and well-conceived policy framework – one based upon incentivising innovation and up-skilling the workforce – rather than generating rents and preserving the status quo. It calls for a European policy based upon and oriented towards the future, and not a sop thrown out casually to the fast-eroding past.

[31] Paul Keller, “What the Diary of Anne Frank Can Tell Us About Text and Data Mining,” *Communia*, 08 January 2016.

[32] LIBER, op. cit.

[33] On 09 December 2015, the European Commission announced that it is considering new copyright laws and proposed a mandatory exception for research in the EU copyright legislation, for “public interest research organisations to carry out text and data mining of content they have lawful access to, with full legal certainty, for scientific research purposes.” See the box on page 2.

[34] On 23 March 2016, the European Commission launched the “Public Consultation on the Role of Publishers in the Copyright Value Chain and on the ‘Panorama Exception.’” Many believe this surprise consultation, which consists of 23 box-ticking questions and little room for commentary or insertion, implies that the European Commission is seeking to extend so-called “neighbouring rights” to the scientific, technical and medical publishing sector as well, effectively over-riding any pledges made elsewhere to offer a research exemption for text and data mining.

‘The weight of activity is shifting towards new markets.’

References and further reading

Cocoru, Diana, and Mirko Boehm. *An Analytical Review of Text and Data Mining Practices and Approaches in Europe. Policy Recommendations in View of the Upcoming Copyright Legislative Proposal* (London: OpenForum Europe, 2016)

European Commission. *Report from the Expert Group on Standardisation in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining* (Brussels: European Commission, 2014)

Filippov, Sergey. *Mapping Text and Data Mining in Academic and Research Communities in Europe* (Brussels: The Lisbon Council, 2014)

Handke, Christian, Lucie Guibault and Joan-Josep Vallbé. “Is Europe Falling Behind in Data Mining? Copyright’s Impact on Data Mining in Academic Research,” *Social Science Research Network*, 20 May 2015

Ligue des Bibliothèques Européennes de Recherche/ Association of European Research Libraries (LIBER). *Text and Data Mining: The Need for Change in Europe* (The Hague: LIBER, 2014)

Organisation for Economic Co-operation and Development. *Data-Driven Innovation: Big Data for Growth and Well-Being* (Paris: OECD Publishing, 2015)

———. “Making Open Science a Reality,” *OECD Science, Technology and Industry Policy Papers, No. 25* (Paris: OECD Publishing, 2015)

Triaille, Jean-Paul, Jérome de Meeûs d’Argenteuil and Amélie de Francquen. *Study on the Legal Framework of Text and Data Mining* (Namur: De Wolf and Partners, 2014)

Tsai, Hsu-Hao. “Global Data Mining: An Empirical Study of Current Trends, Future Forecasts and Technology Diffusions,” *Expert Systems with Applications*, 39(9): 8172–8181, 2012

The Lisbon Council asbl
IPC-Résidence Palace
155 rue de la Loi
1040 Brussels, Belgium
T. +32 2 647 9575
F. +32 2 640 9828
info@lisboncouncil.net
www.lisboncouncil.net
twitter @lisboncouncil

ISSN: 2031-0935 (digital);
2031-0927 (print)

This interactive policy brief is published by the Lisbon Council.
The responsible editor is Paul Hofheinz, president, the Lisbon Council.

Design by karakas