theLisboncouncil
think tank for the 21st century

Mapping Text and Data Mining in Academic and Research Communities in Europe

special briefing

Issue 16/2014

By Sergey Filippov

Sergey Filippov is associate director of the Lisbon Council. A Dutch national, he previously served as assistant professor of innovation management at Delft University of Technology and holds a PhD in economics and policy studies of technical change from UNU-MERIT, a joint research institute of the United Nations University and University of Maastricht.

This report aims to map the scale of the use of text and data mining practices in the academic and research community in several European countries, and to benchmark the state of play in Europe against leading Asian and American countries.[1] The study first provides an overview of academic publications and patents pertaining to text and data mining. This quantitative data is then supplemented by in-depth interviews with leading researchers and text and data mining experts from a number of European countries and the United States. The paper is intended to serve as impartial input into current policy debates on reforming European copyright legislation to make it fit for the challenges and opportunities of the digital age.

Among the study's key findings, which will be analysed and developed in greater detail in the following pages, are the following:

1. Text and data mining is already an important tool for making sense of and finding value in data. The world of data is growing exponentially, offering new insights, better analytics and deeper understanding in a wealth of areas (including human health, analysis of traffic and migratory patterns, climate and environmental systems and more).[2]
2. There has been a strong increase in recent years in the number of publications and patents referring to text and data mining around the world. This growth is driven mostly by US and Asia (China in particular). US nationals are responsible for almost half of all publications and patents in the text and data mining field.

This interactive special briefing 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.

The opinions expressed in this special briefing are those of the author alone and do not necessarily reflect the views of the Lisbon Council or any of its associates.

1
The author would like to thank Christian Reimsbach-Kounatze of the Organisation for Economic Co-operation and Development (OECD), Gaetan De Rassenfosse of the University of Melbourne, and Ian Hargreaves and Paul Hofheinz of the Lisbon Council for comments on an early draft. Special thanks as well to Maja Bogataj Jančič, Antal van den Bosch, Henrik Boström, Alex Coad, Rishab Ghosh, José Guimón, Jun Hou, Can Huang, Pim Huijnen, Ashwin Ittoo, Dinar Kale, Paul Keller, Jan Knoerich, Izaskun Lacunza, Georg Licht, Boris Lokshin, Sarianna M. Lundan, Alex Mohr, Ismael Rafols, Nico Rasters, Francesco Rentocchini, Alexander Settles, Mariagrazia Squicciarini, Dariusz T. Stepniak, Jie Tang, Vlad Vaiman, Tim Vink, Fardad Zand and Erika Zoeller Véras. As always, any errors of fact or judgment are the author's sole responsibility.

2
Paul Hofheinz and Michael Mandel, Bridging the Data Gap: *How Digital Innovation Can Drive Growth and Create Jobs* (Brussels & Washington, DC: Lisbon Council and PPI, 2014).

'Text and data mining is already a key tool for making sense of and finding value in data.'

By contrast, European universities and scholars lag behind (see Table 2 and Table 7 on pages 9 and 10 for more). As new discoveries and advances so often build upon previous findings, this gap could widen even further.

3. Europe's intellectual property regime – including copyright – may have worked well in an economy where knowledge was stored mostly in text on paper and analysed typically by scholars taking notes. In the digital age, 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. Interviews with researchers in Europe suggest that these conditions do not exist (see Annex I and Annex II, which begin on page 24, for a summary of academic interviews conducted for this study).
4. Europe's current weakness in this area could have long-lasting repercussions. If specific text and data mining techniques are protected by patents granted overwhelmingly to third-country nationals, European researchers may face legal obstacles when trying to advance knowledge in this field.
5. Europe's academic community is aware of the growing role of data analytics and text and data mining. In a survey conducted for this special briefing, we found that many academics already use text and data mining at the basic level; and many more are considering using it in the future as the process becomes simpler and the benefits more easily seen.
6. The research community itself is increasingly in favour of so-called "open access" to scientific publications, with an increasing volume of academic literature now published online to the reader free of charge and free of most copyright and licensing restrictions.[3] In turn, many academic publishers acknowledge this growing demand; some are seeking to integrate open access into their existing business models.
7. The use of text and data mining differs among disciplines. Scholars in computer science – where the techniques are most indigenous – use it the most. Scholars in social sciences (economics, management, international business, innovation studies, etc.) still use the methodology only sporadically.

Data: Crude Oil of the Digital Age

The pace of innovation and technical change has been greatly increased by revolutionary advances in information and communication technology (ICT) over the past two decades. With the proliferation of the Internet, the rise of the user-generated content and the digitalisation of traditional industries and services, we have witnessed an explosive growth of data.[4] Businesses collect trillions of bytes of information on customer transactions, suppliers, internal operations and indeed competitors. Social networking sites such as Facebook and Twitter enable users to share over 1.3 billion pieces of information and content per day. Facebook users share over 30 million pieces of content per month, and Twitter has 350 million tweets daily.[5]

This phenomenon has come to be known colloquially as "Big Data." It spans three different dimensions – volume, velocity and variety.[6] In other words, data are generated

3 Open access removes price barriers (such as subscriptions, licensing fees, pay-per-view fees) and permission barriers (copyright and licensing restrictions). Open access journals are financed by academic institutions, academic associations, governments or even academic publishers themselves. While some journals are fully open access, others may grant open access only to certain articles. For more information, see Peter Suber, *Open Access* (Cambridge: MIT Press, 2012).

4 Viktor Mayer-Schönberger and Kenneth Cukier, Big Data: *A Revolution that Will Transform How We Live, Work and Think* (London: John Murray, 2013). See also Sarah Buchholtz, Maciej Bukowski and Aleksander Sniegocki, *Big and Open Data in Europe* (Warsaw: DemosEUROPA, 2014).

5 Diane McDonald and Ursula Kelly, *Value and Benefits of Text Mining* (Bristol: JISC, 2012).

6 Gartner, *Gartner Says Solving "Big Data" Challenge Involves More Than Just Managing Volumes of Data*, Press Release (Stamford: Gartner, 2011).

‘The world of data is growing exponentially, offering new insights, better analytics and deeper understanding.’

at a very fast pace, in huge amounts and from different sources. Presently, humans create 2.5 quintillion bytes of data every day and 90% of the data in today’s world has been created in the last two years. The volume of data is doubling every three years.[7] The staggering pace of growth of digital data can be broadly compared to the pattern of Moore’s Law, which holds that the number of transistors on integrated circuits will double approximately every two years.[8] A key enabling factor in the growth of digital data is the decrease in data storage cost. Other enablers leading to a dramatic surge in data are the deployment of mobile devices and sensors and sensor networks, and the Internet of Things.[9]

No individual can keep up with such a volume of data; scientists need computers to help them make sense of the information. For this purpose, text and data mining software, often combined with artificial intelligence, is being increasingly employed. Data mining 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.[10] 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.

The potential that comes with mining scientific literature is enormous. Text and data mining has enabled insight that otherwise would not be possible to gain. Some remarkable examples include the use of text mining in hypothesis development in conceptual biology,[11] an automated framework for hypotheses generation using literature,[12] automatic detection of adverse events to predict drug label changes,[13] mining the astronomical literature[14] and speed reading.[15]

A report published by McKinsey Global Institute in 2012 suggested a substantial economic potential for text and data mining methods. If US healthcare, for example, were to use data more creatively and effectively to drive efficiency and quality, the study found that the sector could create more than $300 billion [or €215 billion] in value every year.[16] In the developed economies of Europe, government administrators could save more than €100 billion in operational efficiency improvements alone by using data more effectively, not including using data to reduce fraud and errors and boost the collection of tax revenues. In total that would lead to $250 billion [or €180 billion] of potential annual value to Europe’s public sector administration, according to the McKinsey study. Users of services enabled by personal-location data could capture $600 billion [or €430 billion] in consumer surplus.[17]

The use of text and data mining also has important repercussions for academia. Annually, the global academic and research community generates over 1.5 million new scholarly articles;[18] with an estimated 50 million academic articles in circulation as of 2010.[19] Such volumes of material have outstripped the ability of individuals to access, read and analyse these publications. Hence, specifically in the global academic and research community, text and data mining has the potential to enhance very significantly current practices.

7 *Big Data at the Speed of Business* (New York: IBM, 2014).

8 The observation was originally formulated by Gordon Moore, Intel co-founder, in 1965. See Gordon E. Moore, “Cramming More Components onto Integrated Circuits,” *Electronics,* 19: 114–117, 1965. David C. Brock (ed.), *Understanding Moore’s Law: Four Decades of Innovation* (Philadelphia: Chemical Heritage Foundation, 2006).

9 OECD, “Exploring Data-Driven Innovation as a New Source of Growth: Mapping the Policy Issues Raised by Big Data,” *OECD Digital Economy Papers*, No. 222 (Paris: OECD, 2013).

10 Data mining is a broader field than text mining. Data mining can also involve mining in images (including moving images), video and other audiovisual media. These data sources find wide application in domains such as healthcare (medical scans and biomedical data), entertainment (music and movies), gaming industries, surveillance and geographic information systems (spatial data mining).

11 Tanja Bekhuis, “Conceptual Biology, Hypothesis Discovery, and Text Mining: Swanson’s Legacy,” *BioMedical Digital Libraries,* 3(2), 2006.

12 Vida Abedi, Ramin Zand, Mohammed Yeasin and Fazle Elahi Faisal, “An Automated Framework for Hypotheses Generation Using Literature,” *BioData Mining,* 5(13), 2012.

13 Harsha Gurulingappa, Luca Toldo, Abdul Mateen Rajput, Jan A. Kors, Adel Taweel and Yorki Tayrouz, “Automatic Detection of Adverse Events to Predict Drug Label Changes Using Text and Data Mining Techniques,” *Pharmacoepidemiology and Drug Safety,* 22(11): 1189–1194, 2013.

'In the digital age, new technologies make analysis of large volumes of text and other media potentially routine.'

But in many nations, the use of text and data mining clashes directly with the intellectual property regime. The problem is neatly summarised in a position paper of Universities UK and UK Higher Education International Unit:

"Researchers are expected to make sense of vast quantities of in-copyright published material which they have legal access to. However, at the moment, researchers do not have recourse to analytical tools that are taken for granted in many other aspects of modern life, which would assist them in doing so. The extraction of facts or individual words is not subject to copyright law – a human being copying a word or a fact from an article with a pen or pencil is perfectly free to do so. But due to the fact that a computer must make a copy of an entire in-copyright work in order to perform the same activity, the process of data mining becomes subject to copyright law. In short, copyright law which originates from the time of the printing press does not prohibit a human from reading and making associations between the facts contained in content they have bought or have legal access to, but it does prohibit a computer programme from undertaking the same work." [20]

Use of Text and Data Mining: An Analytical Framework

But how much text and data mining is really going on? Who is using it? And where? In this section, we seek to quantify and compare the use of text and data mining in academic circles from different regions. Two methods are employed – a bibliography-based analysis, which looks at the amount of attention text and data mining was given by the academic community as a topic of research, and a patent-based analysis, which counts the number of patents awarded for text and data mining related processes and techniques. The goal is to put concrete figures on the actual use of text and data mining, and to establish its scope and depth of penetration in different areas of the world.

Bibliography analysis

In an effort to quantify references to text and data mining as a topic of scientific publications, we consulted ScienceDirect, a database operated by Elsevier containing about 11 million articles from 2,500 journals and over 25,000 e-books, reference works, book series and handbooks.[21]

The journals available on ScienceDirect are grouped into four main sections: physical sciences and engineering, life sciences, health sciences, and social sciences and humanities. Although our report treats text and data mining jointly, many scientific publications address data mining and text mining separately. As a result, we executed the search for two different terms – "data mining" and "text mining" – in titles of publications. It is important to emphasise that this study looks only at publications that treat text and data mining as their primary subject, i.e. featuring research findings about the advancement of these techniques. Publications that actually use text and data mining as (part of) their methodology are outside the scope of this research.

Only a few publications contained both "text mining" and "data mining" in the same title, the first of such articles was published only in 2009.

14 Alasdair Allan, "Mining the Astronomical Literature," *Radar*, 15 August 2012.

15 Corie Lok, "Literature Mining: Speed Reading," *Nature*, 463: 416–418, 2010.

16 The exchange rate is from March 2014.

17 James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh and Angela Hung Byers, *Big Data: The Next Frontier for Innovation, Competition, and Productivity* (San Francisco: McKinsey Global Institute, 2011).

18 Diane McDonald and Ursula Kelly, *Value and Benefits of Text Mining* (Bristol: JISC, 2012).

19 Arif Jinha, "Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence," *Learned Publishing*, 23(3): 258–263, 2010.

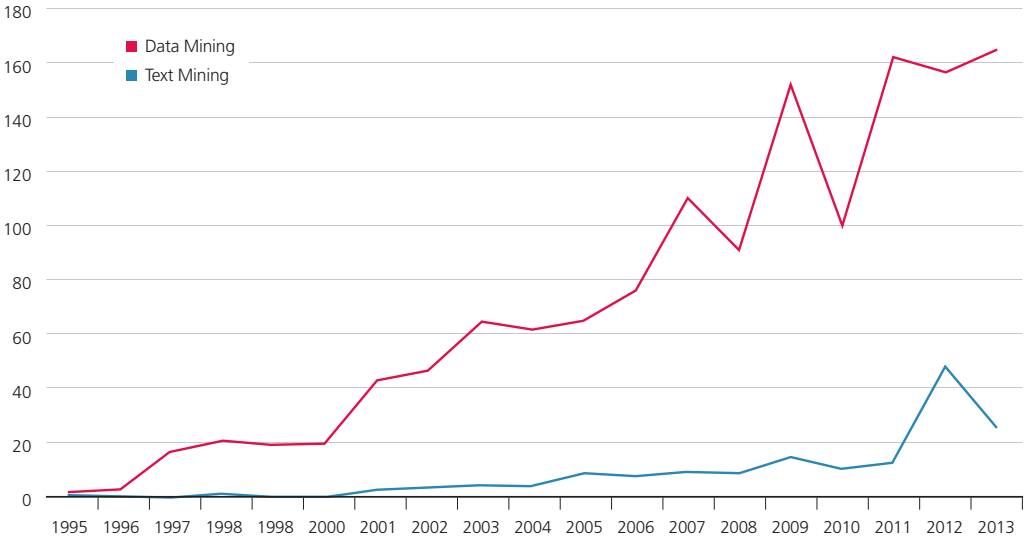
20 Universities UK and UK Higher Education International Unit, *European Commission's Stakeholder Dialogue 'Licenses for Europe' and Text and Data Mining* (London: Universities UK, 2013).

‘Europe’s current weakness in this area could have long-lasting repercussions.’

The results are shown in Chart 1 and Table 1. Two main conclusions can be drawn: 1) data mining as a field of academic inquiry has received more interest than text mining, and 2) there is an evident upward trend in the number of publications in both domains.

Chart 1: Number of Published Academic Articles about Text and Data Mining

In ScienceDirect database, per year



Source: Lisbon Council calculation, using data from ScienceDirect database

Table 1: Articles on Data Mining and Text Mining

In ScienceDirect database, per year

Year	Data mining	Text mining	Both text and data mining
1995	2		
1996	3		
1997	16		
1998	21	1	
1999	19		
2000	19		
2001	43	2	
2002	46	3	
2003	65	4	
2004	62	4	
2005	65	9	
2006	76	7	
2007	110	9	
2008	91	9	
2009	152	15	3
2010	100	10	1
2011	162	12	4
2012	157	48	7
2013	165	25	1
Total	**1374**	**158**	**16**

Source: calculated by author, using data from ScienceDirect database

21
Specifically, we searched the titles of publications – and did not include the full texts or abstracts of the texts in the search criteria (or seek to determine if an article was written using text and data mining techniques). All queries were manually crosschecked to avoid unrelated articles, such as “data on coal mining.” The goal was to determine where text and data mining were being given the most attention within the academic community as a subject of study. We did this by counting the number of academic articles about text or data mining per year (we believe that an article that has text or data mining in the title will be devoted to that topic). The search did not include articles that only mention text or data mining in the main body. And it did not count the number of times text and data mining were mentioned in text bodies in the years surveyed. As a result, the survey gives a conservative account of the actual role of text and data mining as an academic topic in the years covered. www.sciencedirect.com. Abstracts of most articles were freely available for research, but access to the full text of the article required a subscription or pay-per-view purchase.

Text and Data Mining in the News

The controversies surrounding text and data mining practices, particularly in research and academia, have become the subject of several mainstream and popular science-press articles. A selection follows:

Alok Jha, "Text Mining: What Do Publishers Have Against This Hi-Tech Research Tool?," *The Guardian*, 23 May 2012

> "Researchers push for end to publishers' default ban on computer scanning of tens of thousands of papers to find links between genes and diseases… Unfortunately, in most cases, text mining is forbidden. …[C]ountless … academics are prevented from using the most modern research techniques because the big publishing companies such as Macmillan, Wiley and Elsevier, which control the distribution of most of the world's academic literature, by default do not allow text mining of the content that sits behind their expensive paywalls."

Richard Van Noorden, "Trouble at the Text-Mine," *Nature*, 483, pp. 134–135, 2012

> "Computers can rapidly scan through thousands of research papers to make useful connections, but work is being slowed by publishers' unease. When he was a keen young biology graduate student in 2006, Max Haeussler wrote a computer program that would scan, or 'crawl,' plain text and pull out any DNA sequences. To test his invention, the naive text-miner downloaded around 20,000 research papers that his institution had paid to access – and promptly found his IP address blocked by the papers' publisher."

Richard Van Noorden, "Text-Mining Spat Heats Up," *Nature*, 495, p. 295, 2013

> "Scientists and publishers clash over licences that would let machines read research papers. It is seen as the future of computer-based research — if only the gatekeepers would let scientists in. Researchers have complained bitterly over the past year that publishers do not allow them to use computer programs to download and crawl across the text of research articles, a methodology known as text mining that can reveal large-scale patterns in the studies."

Richard Van Noorden, "Tensions Grow as Data-Mining Discussions Fall Apart," *Nature*, 498, pp. 14-15, 2013

> "Scientists want to exempt computer-based text crawling from Europe's copyright law. Disagreement between scientists and publishers has grown on

a thorny issue: how to make it easier for computer programs to extract facts and data from online research papers. On 22 May, researchers, librarians and others pulled out of European Commission talks on how to encourage the techniques, known as text mining and data mining. The withdrawal has effectively ended the contentious discussions, although a formal abandonment can be decided only after a commission review in July."

Mark MacCarthy, "Text Mining Revolutionizes Academic Research," *Digital Discours*, 4 February 2013

"The benefits of big data analytics extend well beyond the uses by businesses and governments. As the following examples illustrate, the version of data analytics known as text mining is an essential part of how the scholarly and scientific community does its research. A well-known research finding, cited in the recent Hathitrust decision, illustrates the benefits of text mining in literary and historical research. By comparing the frequency with which authors used "is" to refer to the United States rather than "are" researchers were able to conclude that it was only in the second half of the 19th Century that we began to think of our nation as a single, indivisible entity."

ScienceGuide, "Victor Hugo Still Rules EU Copyright," *ScienceGuide*, 16 October 2013

"Big data, and the possibility to 'mine' these big data, has huge scientific potential. But in Europe Victor Hugo still rules copyright. That means the right to read is not the right to mine. So why is text and data mining important for digital innovation? Paul Keller: "Text and data mining is a significant tool for scientific research. In South-Korea, Japan and in the U.S. you can use text and data without any copyright-restrictions. In the EU however, you need extra permission to use a newspaper database for research. Another problem with European copyright is that it is completely scattered: all European countries have different rules regarding quotations and the licenses for educational purposes vary broadly."

In essence, all these publications accentuate the advantage of text and data mining and highlight the benefits it can bring to the academic community, and more broadly, to the society. The current legislation is portrayed as outdated and the power and business models of global academic publishers as not corresponding to the realities of the digital age.

'We found that many academics already use text and data mining at the basic level.'

Specifically, among the 11 million articles in the database, we found 1,374 articles devoted to data mining and 158 about text mining. We also found 16 publications that examine the interplay of text mining and data mining.

Not surprisingly, publications devoted to text and data mining emerged only in the late 1990s as the technology developed.[22] Initially, their number was rather modest – ranging from one to 21 articles per year. The number of data mining publications saw a sustained growth to 165 in 2013, up from 43 in 2001, a year which proved to be a turning point. It is worth mentioning that the growth was not smooth, and the spikes in Chart 1 on page 5 can be explained by the nature of academic publishing procedures and the lag between completion of research and its appearance in a journal. Research submitted to a journal may undergo a series of reviews; further, a journal may not have an empty slot for a ready manuscript in a forthcoming issue. It may well be that 2010 was a productive year for the research community but the results are published only in 2011 or even 2012.

Hsu-Hao Tsai, a researcher in the management information systems department at National Chengchi University in Taiwan, conducted a similar exercise in 2012.[23] His paper uses a bibliometric approach, analysing technology trends and forecasts of data mining from 1989 to 2009 by locating the heading "data mining" as the topic in the Social Science Citation Index (SSCI) database.[24] The year 1989 was taken as a starting date as no prior data existed on this topic. In total, 1181 articles were found, which roughly matches our numbers retrieved from ScienceDirect.[25] Generally, Mr Tsai's findings from SSCI corroborate our finding from ScienceDirect – the growing attention paid to data mining in academic publications. Similarly, it is closely followed by a rise in citations of these publications.

Table 2 on page 9 presents the distributon of top countries whose scientists produced the highest volumes of articles on data mining, according to Mr Tsai's study. Clearly, US scientists are in the lead, authoring virtually half of all publications on data mining. The UK ranks second and Taiwan – third. Next to the UK, there are 10 other European Union member states on the list – Austria, Belgium, Finland, France, Germany, Italy, Netherlands, Slovenia, Spain and Sweden. Taken together, publications originating from the 11 EU member states amount to 311, or more than one quarter (26.3%) of the sample.

Beyond the mere fact of appearance in an academic journal, citations are perhaps even more important as they are objective evidence of a study's usability and relevance. Therefore, we set out to measure this by taking Mr Tsai's study as a base and calculating a new variable – "citation per publication" that would enable us to understand the impact of publications. Citations per publication is an average for all publications produced by researchers of a particular country. It might well be the case that only one publication is widely cited while others do not receive any exposure at all. Nevertheless, we use this averaged indicator for the purposes of our analyis. As the data in Table 2 show, this indicator ranges from 0.4 (Slovenia) to 47.4 (Finland), meaning an article published by Finnish researchers was cited on average 47 times, and in Slovenia – roughly speaking – only one out of two articles would be cited.

22 The query was executed with the key words "data mining" and "text mining." The author acknowledges that other terms were sometimes used in the past for what is now known commonly as "text mining" and "data mining," such as "knowledge discovery." A broader search is beyond the scope of the study.

23 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. Mr Tsai's research is focused on text and data mining in the social sciences.

24 Social Science Citation Index (SSCI) database.

25 It is worth noting the study did not consider "text mining" publications.

‘No individual can keep up with such a volume of data. Scientists need computers to help them make sense of the information.’

Table 2: Distribution of Geographic Sources of Published Articles on Data Mining (1989-2009)

	Country	Number	Share in total (%)	Citations	Citations per publication
1	USA	551	46.66	4781	8.7
2	**UK***	**131**	**11.09**	**1159**	**8.8**
3	Taiwan	104	8.81	436	4.2
4	Canada	67	5.67	547	8.2
5	China	54	4.57	187	3.5
6	Australia	47	3.98	350	7.4
7	**Germany**	**32**	**2.71**	**177**	**5.5**
8	South Korea	32	2.71	232	7.3
9	**Spain**	**27**	**2.29**	**79**	**2.9**
10	**The Netherlands**	**21**	**1.78**	**135**	**6.4**
11	**Belgium**	**20**	**1.69**	**96**	**4.8**
12	**France**	**20**	**1.69**	**105**	**5.3**
13	Japan	18	1.52	49	2.7
14	**Italy**	**17**	**1.44**	**78**	**4.6**
15	Brazil	13	1.10	33	2.5
16	South Africa	13	1.10	69	5.3
17	**Sweden**	**12**	**1.02**	**11**	**0.9**
18	Turkey	12	1.02	53	4.4
19	India	11	0.93	30	2.7
20	**Slovenia**	**11**	**0.93**	**4**	**0.4**
21	**Austria**	**10**	**0.85**	**30**	**3.0**
22	**Finland**	**10**	**0.85**	**474**	**47.4**
23	Singapore	10	0.85	105	10.5

Source: Adapted from Tsai (2012). EU member states are shown in bold. The column “citations per publication” was calculated by the author.
*The original article provided separate data for England, Scotland and Wales. Aggregated data for the UK was calculated by summing up inputs from these three constituencies.
Note: The percentage points calculated in Tsai (2012) relate to the total number of publications – 1181. The sum of publications of individual countries (1243) is higher than this number due to joint publications assigned to 2 or more countries. Therefore, share in total in % is above 100% too.

'The potential that comes with mining scientific literature is enormous.'

The mean value for all countries is 6.8 citation per publication, and for EU member states in the sample – 8.2 (this number is greatly inflated by Finland). Among EU leaders are Finland (47.4), UK (8.8) and The Netherlands (6.4); and among the laggards are Slovenia (0.4), Sweden (0.9) and Spain (2.9).

These citation indices should be taken with caution, as it doesn't necessarily tell us much about the quality of the publication as such. Other factors may be of relevance, e.g. the journal (articles in top-tier journals have a higher propensity to be cited) and the language (articles not published in English are often ignored by the global academic community).

Table 3 below presents an overview of languages in which research on data mining is published. Unsurprisingly, English is the leading language. It is used not only by native speakers but also other scientists as it is the de facto lingua franca in academic research. Spanish ranks second (reflecting a combined potential of Spain and Latin American countries), and German is third.

Table 3: Distribution of Articles on Data Mining
By language, from 1989 to 2009

	Number	Share (%)
English	1149	97.29
Spanish	12	1.02
German	5	0.42
Slovak	4	0.34
Japanese	3	0.25
Czech	2	0.17
French	2	0.17
Portuguese	2	0.17
Russian	1	0.08
Slovene	1	0.08
Total	**1181**	**100**

Source: Tsai (2012)

Table 4 on page 11 offers distribution of publication on data mining over research areas. The top three subjects for data mining research are information science and library science (260 articles, or 22% of the total set), followed by computer science and information system (251 articles, or 21%), and operational research and management science (168 articles, or 14%). The statistics reveal that data mining is not limited to computer sciences or information technologies only; it finds wide application in social sciences too, such as business and education.

‘Text and data mining has enabled insight that otherwise would not be possible to gain.’

Table 4: Distribution of Articles on Data Mining by Subjects (1989- 2009)

Rank	Subject areas	Number	Share (%)	Citation
1	Information Science & Library Science	260	22.02	1508
2	Computer Science, Information Systems	251	21.25	1941
3	Operations Research & Management Science	168	14.23	1096
4	Management	149	12.62	864
5	Computer Science, Artificial Intelligence	132	11.18	1421
6	Economics	112	9.48	1373
7	Computer Science, Interdisciplinary Applications	103	8.72	713
8	Public, Environmental & Occupational Health	85	7.20	588
9	Engineering, Electrical & Electronic	82	6.94	742
10	Environmental Studies	68	5.76	367
11	Business	56	4.74	350
12	Geography	52	4.40	348
13	Medical Informatics	49	4.15	239
14	Environmental Sciences	38	3.22	378
15	Social Sciences, Mathematical Methods	35	2.96	1123
16	Ergonomics	34	2.88	146
17	Engineering, Industrial	33	2.79	147
18	Planning & Development	31	2.62	201
19	Education & Educational Research	30	2.54	97
20	Social Sciences, Interdisciplinary	30	2.54	92
21	Sociology	30	2.54	197
22	Mathematics, Interdisciplinary Applications	26	2.20	221
23	Geography, Physical	24	2.03	212
24	Computer Science, Cybernetics	23	1.95	114
25	Statistics & Probability	21	1.78	1207

Source: Tsai (2012)

On the basis of his literature review, Mr Tsai lists areas in which data mining as a research methodology is being actively employed. They are database marketing, interface, semantic indexing, cancer information system, customer retention and insurance claim patterns, data quality, customer service support, electroencephalography application, prediction of corporate failure, network intrusion detection, knowledge refinement, software integration, credit card portfolio management, knowledge warehouse, grid services, library material acquisition budget allocation, selection of insurance sales agents, prediction of physical performance and library decision making.

'Annually, the global academic and research community generates over 1.5 million new scholarly articles.'

As the citation analysis show, subject areas in the top of the list receive the highest number of citations per publication, too. The leader is computer science, information systems with 1,941 citations per publication, closely followed by information science and library science (1,508 citations) and computer science, artificial intelligence (1,421). A very remarkable finding is that the field of statistics and probability has the lowest number of publications on data mining (21), but one of the highest citation scores (1,207 citations per publication). These 21 publications introduce data mining as a methodological approach, and publications in other domains provide citations to them every single time a specific approach is used. The field of social sciences has the lowest number of citations – 92.

The conclusions from our bibliographic analysis in ScienceDirect and the results of Mr Tsai's study in the SSCI database are as follows:

1. Both data mining and text mining are emerging domains of scientific inquiry. The number of text and data mining-related publications sees sustained upward trend, which seems likely to continue, given the growth in data and a growing population of researchers with the necessary skills.
2. The US appears to be leading the field, as virtually half of all publications are produced by scholars affiliated with US universities. The UK is the leader in Europe, and, jointly, scholars affiliated with universities of EU member states are responsible for almost a quarter of all publications. Among nations on the North American continent, Canada has a strong position; and Australia, China, Japan, South Korea and Taiwan are leaders in the Asian region.
3. Information science and library science, computer science and information systems are the fields in which text and data mining-related publications mostly appear (jointly responsible for 43%). However, text and data mining is not solely limited to "hard" sciences. Text and data mining-related articles appear in the field of management, social sciences and education.

Patent analysis

We also set out to gauge the use of text and data mining in academic and research communities in Europe and beyond by counting the number of text and data mining-related patents that may cover text and data mining algorithms, practical application of text and data mining, etc. For this, we relied on the EspaceNet patent database of the European Patent Office, using the worldwide database functionality that allows searching for information about published patents from over 90 patent-granting authorities.[26] In this study, we counted the number of actual patents awarded, and not just the number of applications. And we looked at the patents in text and data mining from three angles:

- distribution by patent granting authority;
- distribution by the inventor's nationality;
- distribution by the owner's nationality (only the patents granted by European Patent Office).

26 For more information, visit the European Patent Office.

‘There were an estimated 50 million academic articles in circulation as of 2010.’

The first step of the patent analysis was to assess the scale of patentability of data mining. We executed a search query in the EspaceNet database; the key words “data mining” used to search for text in the subject or abstract of a patent.[27] The results are shown in Table 5.

27 The International Patent Classification (IPC) was restricted to G06F – “electric digital data processing.”

Table 5: Patents in Data Mining, by the Patent Granting Authority

Data mining in the title or abstract of a patent

	Total	AU	CA	CZ	CN	DE	EP	GB	GR	HK	IL	JP	KR	MX	RU	TW	UA	US	SG	WO
2000	**47**		1				1					9	4					21		11
2001	**93**		2	1		1	1	1				23	9					41		14
2002	**160**						5	2				22	13			2		98	1	17
2003	**182**	3			1	1	2	1				20	15			2		123		14
2004	**147**				4	1	4	1	2			19	8	1		2	2	94		9
2005	**144**	1			5		4					21	2			3		99		9
2006	**127**	2	1		4	1	5	1		1		13	2	1		3		76		17
2007	**135**		3		9							16	6			3		83		15
2008	**147**	1	4		28		3					6	6	1		3		82		13
2009	**192**	1	2		44		2	2			1	7	9	2	1	3		107		11
2010	**178**	3			67		4	3				7	10		1	5		69		9
2011	**207**	3	1		82		2	1				5	14			5		75		19
2012	**220**	1	2		121		3	1				8	6			6		61		11
2013	**300**		2		178		3	3				2	25			3		71		13
Total	**2279**	**15**	**18**	**1**	**543**	**4**	**39**	**16**	**2**	**1**	**1**	**178**	**129**	**5**	**2**	**40**	**2**	**1100**	**1**	**182**

Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office
Country codes: AU – Australia, CA – Canada, CZ – Czech Republic, CN – China, DE – Germany, EP – European Patent Office, GB – United Kingdom, GR – Greece, HK – Hong Kong, China, IL – Israel, JP – Japan, KR – Korea, MX – Mexico, RU – Russia, TW – Taiwan, UA – Ukraine, US – USA, SG – Singapore, WO – patents granted by WIPO (World Intellectual Property Organisation)

In total, 2279 patents in data mining were granted by several national authorities in the 2000-2013 period. A strong upward trend is detected – in 2000 only 47 patents were granted, and in 2013 this number reached 300. Almost half of all these patents (1100) were granted by the United States Patent and Trademark Office. The State Intellectual Property Office of the Peoples’ Republic of China (SIPO) granted 543 patents over the same period. And there is a strong upward trend, too – in 2003, only one patent was granted, and ten years later, in 2013, this number rose to 178.

‘US scientists are in the lead, authoring virtually half of all publications on data mining.’

Over the same period (2000-2013), the European Patent Office granted only 39 patents. Here a caveat, however, should be added due to the nature of patents in this area. Text and data mining is normally embodied in software. And patent applications for software vary per country not only due to legal but also cultural differences. Software is patentable in the US and other jurisdictions. Under the European Patent Convention, and in particular its Article 52, “programmes for computers” are not regarded as inventions for the purpose of granting European patents, however, there are certain exclusions from this rule. As a result of partial exclusions, and despite the fact that the European Patent Office subjects patent applications in this field to a much stricter scrutiny when compared to its US counterpart, that does not mean that all inventions including some software are de jure not patentable.

To sum up, text and data mining tend to be patented as algorithms in the US and other jurisdictions, and in Europe, text and data mining are patented as being embedded in hardware. Patents granted by the European Patent Office often cover “system, method and apparatus.” This is the main explanation of the low number of patents in data mining granted by the European Patent Office.

The next step of the analysis is to narrow down our search and examine the geography of patents in terms of the nationality of inventors. The search query was restricted, looking at patents containing the terms “data mining” or “text mining” in the title (not in the patent abstract), in order to focus only on those patents that relate very closely to text and data mining.

Each patent has an inventor and an applicant (an individual who filed for a patent and owns it). In many cases, the same person may be the inventor and the applicant. In other situations, an inventor may be an employee of a company that finally patents the invention, or a scientist as an inventor and a university as an applicant. In many cases, both inventor and applicant are based in the same country; but not necessarily, e.g. a product can be invented in an overseas subsidiary of a multinational company.

As we were interested in the countries where text and data mining invention actually happened, we looked at the nationality of inventors, not at the patent owners. In cases of groups of inventors with different nationalities, we attribute the patents to multiple countries according to the proportion of the inventors from these countries (a rule of thumb employed in patent analysis). Our sample shows that such co-invention occurs mostly among scientists sharing common culture and languages, e.g. US and British, US and Canadian, US and Indian, Chinese and Taiwanese nationals. However, there are also other examples of inventors, e.g. German and French, US and Polish, etc.

Since data mining and text mining are often seen as two separate research areas, we searched for these two terms separately. We executed queries with the words “text mining” or “data mining” in the title of a patent. In a few cases, patents covered these two fields simultaneously, but overall we could observe a clear distinction between them. The results are presented in Tables 6 and 7 on pages 15 and 16.

‘The number of text and data mining-related publications is seeing a sustained upward trend, which seems likely to continue.’

Table 6: Patents Granted in Text Mining
By inventor’s nationality

Year	Total	America		Asia				Europe			RoW	
		USA	Canada	Japan	China	Korea	India	Germany	UK	France	Israel	Unknown
2000	1	1										
2001	3			2							1	
2002	9	1	1	5	1						1	
2003	8	5		2							1	
2004	12	5		5		1					1	
2005	11	4		6				1				
2006	14	4		9	1							
2007	15	2		9	1	3						
2008	7	4		2						1		
2009	10	1		5	1	1		1			1	
2010	12	3		5	2			1				1
2011	12	2		7	1	1		1				
2012	12	2		4	4		1		1			
2013	9	3		1	2	2		1				
Total	**135**	**37**	**1**	**62**	**13**	**8**	**1**	**5**	**1**	**1**	**5**	**1**

Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office

The first finding is that more patents have been granted in data mining as opposed to text mining – 927 in contrast to 135 (for the 2000-2013 period). This is unsurprising as the field of data mining is more advanced than text mining, and the finding corroborates the evidence already presented on scientific publications (see Table 1 on page 5).

Second, regarding the nationality of inventors, US scientists are undoubtedly leaders in the data mining field, while Japan leads in text mining. China is rising strongly, especially in data mining. As for the European nations, their performance is rather disappointing. Among EU member states, scientists from Germany and the UK are leaders. While US scientists are responsible for almost a half of all patents in data mining, their EU colleagues are only responsible for just under 8%.

‘China is rising strongly, especially in data mining.’

Table 7: Patents Granted in Data Mining

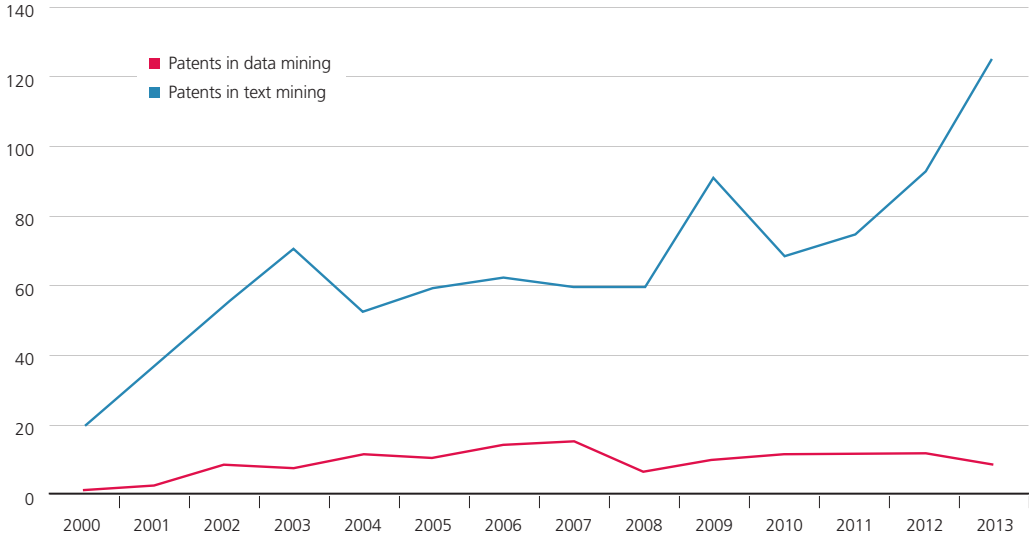
By inventor’s nationality

		2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	**Total**
Rest of the World	Switzerland														1	**1**
	Russia						0.7	1								**2**
	Iceland								0.7		0.5	0.7				**2**
	Australia							2		2.5		1	1			**7**
	Israel		1	2	1	0.8	1		0.3	1	1.8	1	1	1		**12**
Europe	Ireland														0.3	**0**
	Poland													0.5		**1**
	Netherlands														1	**1**
	Italy					1										**1**
	Hungary											1				**1**
	Denmark			1												**1**
	Bulgaria														0.5	**1**
	Greece					2										**2**
	Belgium				1		0.8		0.3							**2**
	Austria				1				0.3		0.3	0.3				**2**
	Sweden										1	0.7			1	**3**
	France	0.5								1	1.3			0.3		**3**
	Spain									1		1	1.2		1	**4**
	UK		1	2.5			0.9	0.8	2.3	0.5	1.2			0.3	6.2	**16**
	Germany		1	1	4.4	1.9	1	1.9	2	2.8	4.3		0.8	3.7	2.7	**28**
Asia	Singapore									1		0.3				**1**
	India				0.3	1.1		2.4		1.5	0.5	0.2	3.3	3.5	1.8	**15**
	Taiwan			1.3	1	1	1	1			1	3	4			**13**
	Korea		2	3	8			1	2	2	3	2	9		7	**41**
	Japan	4	12	4	6	3	7	6	3	4		0.5	1	1		**55**
	China			0.3		2		4	6	1	14	26	26	56	72	**207**
Americas	Argentina														1	**1**
	Brazil													1		**1**
	Canada			1.3	1	2		2	1	3	0.3		2	1		**15**
	US	12.5	14	29.3	40.3	36.2	46.4	38.9	41.2	39.3	60	27.6	24.5	21.8	27.5	**460**
Unknown		3	6	8	6	2	0	1	1	0	1	3	1	1	1	**34**
Total		20	37	54	70	53	59	62	60	60	91	68	75	93	125	**927**

Numbers in the final row are rounded off.
Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office.

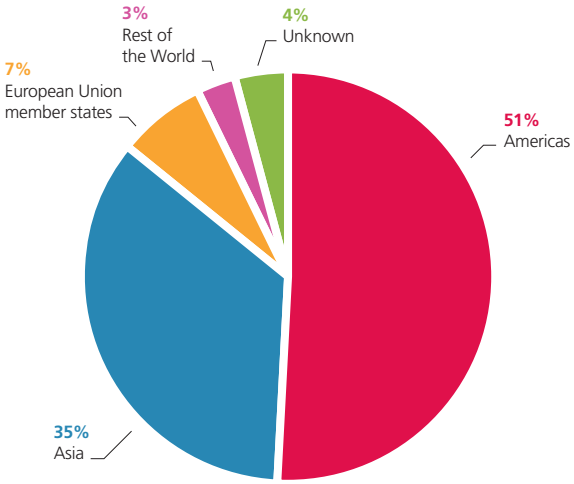
‘Research and academic institutions are important actors in innovation ecosystems.’

Chart 2: Patents in Data Mining and Text Mining Granted in 2000-2013
By inventor’s nationality



Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office

Chart 3: Patents in Data Mining Per Region Granted in 2000-2013
By inventor’s nationality



Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office

As acknowledged earlier in this report, Article 52 of the European Patent Convention does not regard “programmes for computers” as inventions for the purpose of granting European patents, but this exclusion from patentability only applies to the extent to which a European patent application or European patent relates to a computer programme as such. In order to assess the scale of patentability of data mining in Europe, we focused on patents granted by the European Patent Office in this domain.

We executed a search query in the EspaceNet database, but only in the section that contains the entire collection of European patent applications published by the European Patent Office. The key words “data mining” were used to search for text in

'Securing access to a database for text and data mining can be a tedious exercise.'

the subject or abstract of a patent.[28] This time, we looked at the patent applicant, not the inventor, in order to understand who owns patented data mining methods in Europe. The results are given in Table 8 below.

28 The International Patent Classification (IPC) was restricted to G06F – "electric digital data processing."

Table 8: Patents in Data Mining Granted by the European Patent Office

Year	Number
2000	2
2001	3
2002	11
2003	4
2004	15
2005	9
2006	10
2007	1
2008	7
2009	2
2010	4
2011	5
2012	3
2013	4
Total	**80**

Patent Applicant's Nationality	Number
US	56
Germany	5
Korea	4
Japan	3
UK	3
Canada	2
India	2
Belgium	1
Gibraltar, UK	1
Hungary	1
Iceland	1
Total	**80**

Source: Lisbon Council calculation, using data from EspaceNet database of the European Patent Office
Nationality is understood as a registered permanent address of the applicant stated in the patent

In total, 80 patents are identified in the EP Database over the period of 2000-2013. In recent years, around four patents are granted in this domain annually. The highest number of patents in data mining is observed in 2004 (15), and the lowest in 2007 (1). Remarkably, US applicants hold leadership as 56 patents out of 80 are owned by US entities. These are mostly large multinational companies, such as IBM, Microsoft, Siemens and Xerox. Likewise, multinational companies headquartered in other countries are responsible for patent applications in data mining – SAP AG (Germany), LG (Korea), Hitachi (Japan), Philips (The Netherlands), Tata Consultancy (India) and Janssen Pharmaceuticals (Belgium).

Intellectual Property Rights in the Digital Age

Intellectual property is a well-established field of research and its thorough analysis is beyond the scope of this report. We shall only highlight some topical issues here relevant for our discussion of text and data mining.

The protection of intellectual property is vital to IP-intensive industries. A recent study by the European Patent Office and the Office for Harmonization in the Internal Market (OHIM) shows that about 39% of total economic activity in the European Union,

'"Text and data mining is a toolkit that is increasingly easy to use, but it remains a fragmented set of tools," said one researcher.'

worth some €4.7 trillion, is generated by IP-intensive industries.[29] These industries directly provide around 26% of all employment in the EU, according to the study.

The key motivation behind intellectual property regimes is to incentivise inventors by granting them exclusive right to the commercial use of their invention (patents). Likewise, copyright was established to incentivise creation of works of art. Copyright grants the creator of an original work exclusive rights to its use and distribution with the intention of enabling the creator to receive compensation for their work. However, the protection is not absolute. The "fair use" doctrine, enshrined in the 1976 Copyright Act of the US, permits limited copying and distribution without permission of the copyright holder or payment; e.g. commentary, search engines, criticism, news reporting, research, teaching, library archiving and scholarship. Copyright laws in EU member states traditionally provide for "closed lists" of limitations and exceptions that enumerate uses of works that are permitted. Examples of such uses are: quotation, private copying, library privileges, and uses by the media. More often than not these exceptions are highly detailed and connected to specific states of technology, and therefore easily overtaken by emerging new techologies.[30] The take-up of these exceptions also varies from one EU member state to another.

Academic Research and Current Publishing Practice

Research and academic institutions are important actors in innovation ecosystems. They are major producers of scientific knowledge that finds its application in a broad range of economic sectors. The results of such research are presented in the form of peer-reviewed articles published in international scientific journals. While these journals are managed by the international research community itself, global academic publishers provide support through the physical printing of journals and by maintaining electronic platforms that make soft copies of articles available. As a rule, the process of submission, review and publications is free of charge for individual authors; in return, academic publishers retain copyright on respective scientific articles. Other scholars or academic institutions need to pay subscription fees or pay per view in order to access the material.[31]

This model is increasingly challenged by the rising popularity of "open access" publishing. Global academic publishers Taylor & Francis and Routledge, for example, maintain a number of pure open access journals, with no subscription content. The articles in these journals receive both rigorous peer review and expedited online publication. Authors have the option of publishing their open access article under a Creative Commons Attribution (CC-BY) license, as mandated by some research funders.[32]

While open access is a relatively new trend, the classic business model of academic publishers is based on paid access to their databases, either after single payment of a subscription or licence fee. Even then, the terms of the subscription and licence may prohibit reformulation or modification of the copyrighted material. Academic publishers see a potential commercial threat to their primary business from publications arising

29
The European Patent Office and the Office for Harmonization in the Internal Market, *Intellectual Property Rights Intensive Industries: Contribution to Economic Performance and Employment in the European Union.* Industry-Level Analysis Report (EPO and OHIM, 2013).

30
Idem.

31
There are, however, worrisome signs that the current system is breaking down, with growing mistrust between authors and publishers in the academic field. Bernard Forgues and Sébastien Liarte, "Academic Publishing: Past and Future," *M@n@gement*, 16(5): 739–756, 2013; The Economist, "No Peeking… A Publishing Giant Goes After the Authors of Its Journals' Papers," *The Economist*, 11 January 2014.

32
Taylor & Francis Open and Routledge Open

'"Access is more than just being able to download something," said one researcher.'

from the use of text and data mining, known as "derivative works." As the user has to download these databases on his or her computer to conduct text or data mining, publishers are worried about unauthorised access to the data and potential copyright infringements and often strenuously object to legislative changes that would make this routine copying for data-mining purposes legal. However, many representatives of the research academic and research community argue that they should enjoy all rights to analyse articles in any possible form once access has been granted on legitimate grounds.

The UK Publishers Association says that data mining is already widely accepted and practised within the journal publishing arena, without the need for sweeping legislative changes. This claim was supported by *Journal Article Mining: A Research Study into Practices, Policies, Plans…and Promises*, a study commissioned by the association.[33] Methodologically, it is a self-report study, in which representatives of publishing companies answered a set of formulated questions. It reports that over 90% of publisher respondents already granted research-focused mining requests, across academic and professional publications. 32% of the publishers allowed for all forms of mining without permission (while 85% of those 32% are Open Access publishers, not subscription-access publishers). The research also found that "publishers tend to treat mining requests from third parties in a liberal way, certainly so for mining requests with a research purpose."

At the same time, 53% of publisher respondents would decline mining requests if the results can replace or compete with their own products or services.[34] This is perhaps expected, but is still a rather remarkable finding. Academic publishers, unlike other providers of data, be it academic articles or other types of content, say they are unwilling to share it with those customers who will use it to generate (and appropriate) additional value in new products.

However, the opinion of libraries and end-users is not so sanguine. UK PubMed Central (UKPMC) is a free digital repository of biomedical and life sciences journal literature published in the UK.[35] Along with open access articles, UKPMC offers licensed content provided by academic publishers. UKPMC set out to create a database of articles on malaria, and found it had to contact 75 publishers to secure this permission – the number of publishers that had published articles with the word "malaria" in the title. The British Library says it takes an average of 16 months to negotiate the necessary copyright permissions with each publisher.[36]

ReedElsevier is also a Publishers Association member. Its rules in relation to text and data mining are stated explicitly on the company website.[37] The publisher seeks to advance values of open innovation and accessibility of research results. Elsevier's policy for subscription content is as follows:

"Academic subscribers: Researchers at these institutions can text mine subscribed content on ScienceDirect for non-commercial purposes, via the ScienceDirect API's. Access is granted to faculty, researchers, staff and students at the subscribing institution. The terms and conditions of this access are summarized below. Please speak to your Elsevier Account Manager for further details. Corporate and other subscribers: Please contact your Elsevier Account Manager to find out more about our text mining products and services.

33 Eefke Smit and Maurits van de Graaf, *Journal Article Mining, A Research Study into Practices, Policies, Plans..... and Promises* (Amsterdam: Publishing Research Consortium, 2011).

34 The Publishers Association, *The Publishers Association Says Hargreaves Data Mining Proposals Are an "Unwarranted Blunt Instrument"*. Press Release (London: The Publishers Association, 2011).

35 The example is provided by Benjamin White, head of intellectual property at the British Library.

36 Benjamin White, *The Right to Read is the Right to Mine – Text and Data Mining for Libraries and Your Users* (London: British Library, 2013).

37 See Elsevier's Content Mining Policies

'"A 'sharing for research purposes' exception in copyright law would help text and data mining tremendously," said one professor.'

Non-subscriber access: Requests are considered on a case-by-case basis, and can be made by emailing UniversalAccess@elsevier.com.
For open access content: Text and Data mining permission are determined by the author's choice of user license. This information is detailed in the individual articles. Read more on user licenses."

Practically, though, access to data for text and data mining is not granted automatically. Each request is considered on an individual basis with a respective Elsevier Account Manager or another company representative.

As part of our study, we sent a request to Elsevier to provide any statistical evidence on the requests and use of text mining techniques by subscribers. On 29 January 2014, we received the following response from Mr Hop Wechsler, permissions helpdesk manager, global rights department:

"You can try performing a keyword search for the relevant terms (e.g."text mining") on ScienceDirect (http://www.sciencedirect.com) or Scopus (http://www.scopus.com). Elsevier doesn't track these metrics separately as far as I am aware."

Springer, meanwhile, states its commitment to open access and text and data mining of open access content on its website. Special software for effective text mining of its articles (including sample code for developers) is provided.[38]

"As of 14 May 2014 BioMed Central (with Chemistry Central and SpringerOpen) has published 200015 articles of peer-reviewed research, all of which are covered by our open access license agreement which allows free distribution and re-use of the full-text article, including the highly structured XML version. As a result, SpringerOpen's open access corpus is ideally suited for use by text mining researchers."

The International Association of Scientific, Technical & Medical Publishers (STM Association) seeks to standardise the existing practices and to develop common rules for text and data mining:

"As a rightsholder, the publisher must give permission for text mining. This can be done in a number of ways. Permission can be included in an access licence agreement with, for instance, an institution. The International STM Association have produced a model clause for this purpose. Some publishers have established a process for individual researchers to obtain permission to text mine with some restrictions, while others do not support text mining yet. Some organisations such as PubMed, allow unrestricted text mining without permission, although note that this applies to the abstracts only."[39]

As text and data mining practices are burgeoning, universities need to determine their position towards this research method. We searched the websites of major European and American universities but failed to find any example of clear guidelines on the use of text and data mining.

38
Springer, *Using SpringerOpen's Open Access Full-Text Corpus for Text Mining Research.*

39
International Association of Scientific, Technical & Medical Publishers, Text and Data Mining: STM Statement & Sample Licence (Oxford: International Association of STM Publishers, 2012).

'Historians are just learning to collaborate with computational linguists.'

There are cases where UK universities experienced server access being suspended automatically when abnormal access to publishers' databases was being detected.[40] On one occasion, the entire University of Cambridge was denied access to Elsevier's website because a scientist downloaded several dozen papers at once from journals to which the university had already paid subscription fees.[41]

Henceforth, considering the widespread ambiguity surrounding the legal grounds of the use of text and data mining, there is an urgent need for standards and guidelines in this area, both for publishers and research/academic institutions.

40 Benjamin White, *The Right to Read is the Right to Mine – Text and Data Mining for Libraries and Your Users* (London: British Library, 2013).

41 Alok Jha, *Scientists Ask Publishers to Allow Text-Mining Tool That Would Improve Research* (The Raw Story, 2012).

Practical Evidence on the Use of Text and Data Mining in Academia

In order to complement our findings from secondary data sources presented in the first part of this paper, we conducted a series of semi-structured interviews with academic community representatives. Some of the interviews were conducted on the telephone, and in other cases respondents were asked to write down their answers and send back to us by email.

We interviewed two categories of people:

1. Academics (university assistant/associate and full professors), mainly in the field of social sciences, such as international business, management, economics and innovation studies. They may not necessarily be familiar with text and data mining methodology. Our intention here was to understand to what extent text and data mining is widespread as a research method, to inquire about the prospects for text and data mining. For the purposes of our study, we refer to these respondents as "non-experts."
2. Experts in text and data mining. These people may hold academic affiliation (or held it in the past), but more importantly, they specialise in the field of open access, intellectual property rights, text and data mining, etc. We can rely on their extensive expertise to inform our understanding of the subject under investigation. These respondents are further referred to as "experts."

Sampling and data collection

To create a relevant sample, we focused on countries that are leading in text and data mining research. Some 30 invitations were sent to scholars with an invitation to participate in the survey. We received a number of responses; and the range of responses received is given in Table 9 below.

Table 9: Sample of Respondents

	Non-Expert Academics	Text and Data Mining Experts
Europe	UK (5), Germany (2), Spain (2), The Netherlands (2)	The Netherlands (3), Belgium (1), Sweden (1), Slovenia (1)
America	USA (3)	

‘European researchers may face legal obstacles when trying to advance knowledge in this field.’

Semi-structured interviews

We approached all academics with the following set of questions:

1. How widespread is the use of text and data mining in your research community?
2. How widespread is the use of text and data mining in your own research practices?
3. To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?
4. What is your judgment on text and data mining as a research technique?

All interview results are presented in Annex I, which begins on page 24. At the same time, text and data mining experts and data managers were given some flexibility. They could either follow this structure, or provide their reflections in a broader manner. The responses of the experts are included in Annex II, which begins on page 28.

Summary and Conclusions

It is a common practice that academic publishers restrict access to their databases of academic material, particularly for text and data mining purposes. While they state their commitment to open access and academic freedom, securing access to a database for text and data mining can be a tedious exercise. In many instances, requests are handled on a case by case basis, involving several account managers of the publishing company. The resulting product (articles with text and data mining results) will be subject to copyright and licensing too.

Analysis of objective indicators – academic publications and patents in text and data mining – reveals several worrying facts for Europe. US nationals are responsible for almost half of all publications and patents in the field. More recently, there are signs of strong performance from Asian countries (Japan, Korea, China) with regard to both publications and patents. Europe’s current weakness in the text and data mining domain may have long-lasting repercussions. If specific text and data mining techniques are protected by patents granted to third-country nationals, European researchers may face legal obstacles when trying to advance knowledge in this field.

Generally, virtually all respondents surveyed in our study have shown at least minimal knowledge of data analytics and text and data mining. As for the use of text and data mining, the results are mixed. While some respondents actively employ this methodology, others are not yet that technically-savvy, and perhaps not trained to employ text and data mining. Scholars in social sciences (economics, management, international business, innovation studies, etc.) use text and data mining only sporadically, as text and data mining is not yet a mainstream research method. At the same time, the potential of text and data mining is acknowledged by most researchers. Respondents see the benefits of using automated tools to mine the literature and available research data.

This report is an exploratory study based on both objective evidence (publications and patents) and personal opinions of a group of scholars and researchers. The results indicate that the subject of this study is highly relevant and offer directions for further research. For example, a complementary study can be based on a larger scale examination of the actual use of text and data mining among scientists.

Annex I: Summary of Interview Results

	How widespread is the use of text and data mining in your research community?	How widespread is the use of text and data mining in your own research practices?	What is your judgment on text and data mining as a research technique?	To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?
Professor at Business School, **UK**	Not very, but it seems to be gaining in importance, in particular data mining	I'd like to use it more, but I don't have the time to learn more about it at the moment	It's probably very useful and will become more important	I don't have strong views on this and there are good arguments for and against open access
Lecturer in Innovation and International Development, **UK**	I haven't come across text and data mining practices in my research or my research community. I have some friends in artificial intelligence + semantic web community and those guys are heavily involved in that. Based on interaction with them I will say that it is a quite strong technique for doing research but quality of analysis depends on validity of database or data. They usually find nature of data available to do this analysis is quite limited and not much useful.			
Lecturer in the Economy of China, **UK**	Not very widespread. I am not familiar with someone using this technique.	I have not used this technique to date.	I don't know this method well. It should have the potential to yield important new findings especially due to the masses of information being gathered and processed.	This is a complicated subject – it can be both helpful and problematic, depending on who you ask (whether the author, publisher or reader).
Lecturer in Strategy and Innovation, **UK**		As for myself, I can say that I usually have someone else doing text and data mining on my behalf. When I was in Spain, we had a skilful technician writing queries in php or java to retrieve data from the Internet. I recently moved to the UK and I have not still investigated whether there is the same type of resource available here.	Overall, I believe that having the possibility to access big data via text and data mining will become increasingly important for the success of research projects in the next years. The most reputable academics are heavily relying on that right now and an increasing number of researchers will do the same in the near future.	As for legal protection, my personal opinion is that we should try to overcome such an impediment. The main problem remains that highly rated scientific journals do not have such an open provision, this will hamper the diffusion of open access. Only in the long run we will be able to create reputable open access journals.
Research officer, **UK**	Text and data mining is a major application in my research field: Microeconomics of innovation	It is used intensively in my daily research work. For statistical analysis, regression analyses etc.	Based on what resources obtained in hand, it is in general a reliable way to produce consistent results	Indifference to me. University and research institutions will take care of this issue.

	How widespread is the use of text and data mining in your research community?	How widespread is the use of text and data mining in your own research practices?	What is your judgment on text and data mining as a research technique?	To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?
Professor, **Germany**	I think that it these are not a widespread tools in the econometric innovation research community.	We use text mining technics since a variety of years to infer the industry (if the industry's definition is beyond what one can find in usual SIC/NACE-codes). First time we used it was more than 10 years ago... The most recent study we did using text and data mining was in 2013 to identify certain firms. Similarly, we used technics to identify firms in trouble.	More widespread use will be developed. A problem is that a lot of tests are needed to assess the quality of the results. Here manual inspection of a certain sample is often needed. And in addition, not clear rules are available to formally test for the quality of the outcome. (to my knowledge but I'm not an expert)	Unfortunately, I'd like to say yes to both options. The answer is bit more complicated than a simple "yes" or "no"
Professor in International Management and Governance, **Germany**	It is not common in the papers I read or review	I have not used it	It can reveal potentially interesting things, but requires the right questions to be asked.	Problematic. I am also not a big fan of patenting.
Associate Professor in Economics and Business Administration, **Spain**	My colleges do not use it much, I've only met a couple of researchers using these methods, mainly as a tool to support qualitative research methods, such as interviews, and organize categories. I also met a brilliant professor at MIT who is using "big data" to measure inflation, with great results for countries that are manipulating their stats like Argentina.	Nothing at all	I think it has a great future potential	I find it problematic to the extent that it acts as a barrier to the diffusion of research, and also it may interfere with text mining techniques

	How widespread is the use of text and data mining in your research community?	How widespread is the use of text and data mining in your own research practices?	What is your judgment on text and data mining as a research technique?	To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?
Researcher at a Polytechnic University, **Spain**	In my community in "Innovation Studies" is very common, but it is generally based on structured databases, such as WoS, Scopus or patent databases.	I do bibliometrics, so I do this all the time.	I think the insights are much poorer than people assume. I think the insights gained are very often misleading. And many researchers do not have the qualitative understanding of the area under analysis to properly judge the error of the study. I believe data mining should not be used without complementary qualitative studies.	I do not see protection of academia as a problem. For me the problem is the copyright issues associated with the large databases!!
Assistant Professor in Technology, **The Netherlands**	I haven't been involved in any text and data mining research myself as I was mainly using corporate data either through longitudinal surveys or through statistical office data bases (and thus their confidentiality T&C's apply and no worries about copyright, etc.). Yet, I know there are algorithms and mechanisms exist to break big chucks of communication (e.g. an article) into smaller pieces and mask them down to the word-level and still apply data mining and statistical analyses of interest with good rigor Although this approach may resolve the anonymity and privacy issues, not sure if it can also be an answer copyright/IP concern.			
Assistant professor for Productivity and Innovation, **The Netherlands**	The use is very limited. I believe it is not very useful for the type of research I and many of my colleagues are engaged in. We are mostly interested in firm-level information, sometimes in information on teams of individuals. This sort of information is derivable from databases or surveys. I know only of one recent PhD student here at our department, who wrote his piece on contractual aspects of R&D alliances. For this he read close to 300 R&D contracts and searched for specific word combinations. Perhaps data mining techniques would have been helpful for him.			No specific opinion
Instructor in Management and Global Business, **USA**	In International Business and Strategy I see that there is limited use of data mining in research. At my institution in Computer Science there is interest in this issue. There is this project Big Data Alliance - Big Deal for Economic Growth in New Jersey that attempts to bring together people on this issue and sponsor research. There is interest in tapping into this new market and to prepare students for this work	I do not use data mining.	There is possibility but I have not read that many good articles using this technique in my research area. A lot of talk but not a lot of action.	I find legal protection of academic research helpful.

	How widespread is the use of text and data mining in your research community?	How widespread is the use of text and data mining in your own research practices?	What is your judgment on text and data mining as a research technique?	To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?
Professor of International Management, **USA**	Not widespread at all	Not at all	I think it is a legitimate technique, which may be considered somewhat controversial.	I think there are two sides to this story. One can certainly understand publishers that do not want to lose both control over content and revenue but on the other hand, some researchers may argue that scientific inquiry should not be inhibited by artificial barriers. I think that legal protection of academic research is rather helpful.
Postdoctoral Fellow, **USA**	In biomedical sciences term "data mining" is being used in regard to complex analysis of large sets of data generated by methods such as: high-throughput sequencing, ChIP-on-chip, microarray etc. Since these analyses are getting more and more complex there is a whole discipline emerging that is commonly referred to as bioinformatics. What most bioinformaticians do is analyze tons of data and try to find some useful information in them (produce results). And recently this field has been growing tremendously, mostly because DNA sequencing became cheap and accessible. And the availability of personal genome sequences for scientific research already poses some ethical and legal challenges.	Personally, I have not used these kinds of analysis.	Unfortunately, I do not know much about text mining.	What regards copyright of academic research I do believe that since scientific research is mostly sponsored with public funds, after publishing, it should become a public domain and be available to all interested

Annex II: Interview Results – Who's Using Text and Data Mining in Universities and Academia? Response from Experts

Mr. Nico Rasters, former data manager at Maastricht University (The Netherlands), founder and executive director of DAIGU (Academic Services & Data Stewardship):

"I am not aware of any text and data mining policy at Maastricht University, but that does not mean there are no initiatives. One department is looking into Big Data and the university is hosting a Master class on research data management in April 2014.

First and foremost, it would be great if all scientific literature became available to everyone for free. Though "access" is more than just being able to download something; it also entails the ability to find the right articles (this requires proper metadata, a search engine and a search strategy), and the ability to understand and process the data. Text and data mining definitely could play a role there, and I consider it a valid research technique. Just keep in mind that data is never perfect.

Is text and data mining really used to keep track of academic theories from and end-user point of view? I doubt it goes beyond Google Scholar.

On the other hand, publishers could argue that they are the ones who are helping scientists to keep track of recent developments. After all, they only publish the "best" papers and each journal covers specific fields.

While that may be conceptually true, I've always felt that the publish-or-perish movement has corrupted the system so I'm much more in favour of open archiving. I assume that publishers are worried that their data are being redistributed (which affects their profits), and that worry increases with the amount of data that has been downloaded. But there is no way around downloading all data. Products such as Web of Science and Scopus have rather poor user interfaces that only allow for the most basic queries and output (e.g. frequency tables)."

Prof. Dr. Antal van den Bosch, professor at center for language studies, Radboud University Nijmegen (The Netherlands)

- **How widespread is the use of text and data mining in your research community?**

  "I estimate that in the field of computational linguistics (or (human) language technology, natural language processing), text and data mining accounts for about 25%-30% of all research projects. Most of this work is considered "applied" and is often interdisciplinary in nature, the application domain representing the other discipline."

- **How widespread is the use of text and data mining in your own research practices?**

  "With about half of my projects of the past 10 years having a direct text and data mining component, the use of text and data mining is rather widespread and daily."

- **What is your judgment on text and data mining as a research technique?**

  "Text and data mining is a toolkit that is increasingly easy to use, but it remains a fragmented set of tools without clear architectural meta-knowledge or theory. Reasons for using a tool are mostly heuristic and by rule of thumb."

- **To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?**

  "Academic research should in my view be open. License forms such as Creative Commons for texts, and Open Source licenses for software are vital to ensure this openness, and should be used wherever possible. Copyright on academic research may imply limited access to this research, and this should be avoided.

  In text and data mining the ground material is often copyrighted text, so copyright law prohibits us from republishing this data, which hinders reproduction of experiments; this is not good for the generic scientific standards of text and data mining research. Owners of the rights, mostly publishers, usually reject requests to republish the data for research purposes. In cases in which they allow the redistribution, such redistributable data can become hugely popular and become benchmarks for reproduction studies (e.g. as data being used in 'shared tasks', scientific competitions).

  Sometimes we publish open annotations of the data without the data itself, assuming that others will be able to combine the open annotations with a licensed version of the text.

  It is unclear to what extent we are allowed to publish derivatives (e.g. classifiers) derived from (trained on) copyrighted data. A derivative is not a copy, but may contain traces of the original data (e.g. words or multi-word strings) that allow the corpus to be identified (but not reproduced).

  If a text is released in the public domain or under open licenses, the data is usually welcomed by the research community. Examples are historical copyright-free data, translated documents of the EU (e.g. proceedings, law), or crowd-contributed text collections (e.g. Wikipedia, OpenSubtitles.org), the latter usually released under an open license in which individual authors have waived their copyright.
  If in some way a 'sharing for research purposes' exception could find its way in copyright law, so that copyrighted texts may in fact be shared for research, that would help text and data mining research tremendously, allowing for better reproducibility and faster progress. This does not only go for text and data mining research, but also communication sciences, linguistics, literature, and other fields in the Humanities."

Dr. Pim Huijnen, researcher at the department of history and art history, Utrecht University, The Netherlands

- **How widespread is the use of text and data mining in your research community?**

  "Text and data mining is not widely used in historical research. Nevertheless, historical scholarship has a decade long tradition with quantitative methodologies like database analysis, geographic information systems or network analysis. The speed and ease with which large networks can be digitally arranged, or spatial references can be plotted on digital maps today make that a much wider variety of historians adopt these technologies than the traditional group of experts in computational history. However, text and data mining as a means to grasping 'big data' with the help of statistical and/or linguistic methods is fairly new for historians. There are two reasons for this: 1) Text and data mining is itself a rapidly evolving field of scholarly research. Historians are just learning to collaborate with computational linguists or information retrieval experts to adopt their knowledge to the needs of historical research. 2) Historians hardly work with born-digital data. The literature or textual sources they commonly base their studies on, need to be digitised first. This is a highly extensive task, given the troubles, for example, to correctly OCR older (let alone handwritten) texts. As a result, historians are hardly ever capable of doing 'big data' research in the definition of Mayer-Schönberger and Cukier: that the (text and data mining) analysis of the data as a whole reveals information that the close reading of the individual texts it consists of would never be able to."

- **How widespread is the use of text and data mining in your own research practices?**

  "The Dutch National Library's digitised newspaper archive is one of the historical corpora that comes closest to what could be considered 'big data'. The dataset comprises of almost 100 million articles that span the last four centuries. I am part of a research team of historians that tries to develop text and data mining techniques with the aim of grasping this massive amount of information in innovative ways. We do this in close cooperation with the Information and Language Processing Systems group of Prof. Maarten de Rijke at the University of Amsterdam. Together, we are building a text and data mining tool that is tailored to this particular dataset, as well as to the needs of historians. The research team consists of three PhD students and three postdocs (of which I am one) in history at Utrecht University and the UvA, as well as a programmer and a PhD student in information retrieval. (For more information, see: www.translantis.nl.) We, as historical researcher, have the explicit assignment not to use text and data mining as a way of cherry picking nice anecdotes to interlace our studies with. Contrarily, we are to find approaches in which we can use text and data mining as the foundation of our research. I am optimistic that this is possible. However, as I have only been a couple of months underway, I cannot yet show any results of this undertaking."

- **What is your judgment on text and data mining as a research technique?**

  "In our research group, we are confronted with a tension that the use of text and data mining tools for historical research produces. The greatest advantages of text and data mining lie in its ability of pre-processing: it normalises, lemmatises, formalises or standardises (however you want to call it) the data to make it fit for analysis - research steps that would have been too exhaustive (and tedious) to do by hand. Every one of these steps, however, are interpretative steps. They inevitably steer the analysis (or as we are used to call it: interpretation) in certain direction. Historians, as other humanities scholars, are highly allergic to biases. They heavily rely on hermeneutics: their capacity to make interpretative judgments based on domain knowledge and the exclusion of any biases (of hindsight, preferences, etc.). Obviously, this is never fully possible. However, traditionally, historians usually were aware of their biases. Now, these are locked in black boxes that spit out pieces of information that normal historians are incapable of assessing according to the rules of their discipline. As a result, we use text and data mining techniques that imply little pre-processing as possible. Our tool is capable of generating simple word clouds (based on bag-of-words techniques) and timelines. However, those types of visualisations have real added value - not as analytical tools, but as heuristic techniques. They do the opposite of formalising in the sense of extracting all exceptions. They, contrarily, widen our gaze. Word clouds enable us to make associations or trigger us to look in certain ways that traditional research would never be able to. After all, traditionally we start to make strict selections of our source material. Now, we start the exploration of our research topic with - theoretically - all sources available. As an explorative search method, text and data mining is able to enrich our grasp of our historical material in valuable (and often unexpected) ways."

- **To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?**

  "Copyright law severely hampers our research. The fact that we cannot possess over 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'. To adequately do pattern searching or adopt other forms of diachronic text and data mining, our dataset needs a basic level of completeness or, at least, inner consistency. The digitised newspaper archive, however, lacks by far most of the post-war volumes of our leading newspapers. The National Library desperately tries, but repeatedly fails to convince the publishers of giving up on their copyright – including for research purposes. This makes every attempt to use text and data mining to make diachronic analyses useless from the very start. It is paradoxical that the Library is allowed to present older volumes of newspapers to its visitors in paper, but not on the screen. It really is a pity.

I have experienced equal problems in my previous research project. In my CLARIN-funded pilot project BILAND (www.biland.nl) I aimed at comparing Dutch and German public debates with the use of text and data mining techniques. Whereas the Dutch National Library open-mindedly supports scholarly research by putting their data at our disposal, I had the greatest trouble finding a similar dataset in Germany. There I experienced a much heavier reluctance to give their data away because of copyright law. In general, copyright issues are a returning point of discussion and concern at conferences and meetings on Digital Humanities or 'digital history'."

Dr. Ashwin Ittoo, assistant professor of management information systems, HEC Management School, University of Liege, Belgium

- **How widespread is the use of text and data mining in your research community?**

"I work in this area and consequently, I would say that within my field, there is a lot of development in text and data mining algorithms.

This is mainly from the computer science side, which is concerned with furthering and improving the accuracy and general performance of text and data mining techniques for more sophisticated tasks. For example, there is now an emerging body (emerging is a bit strong... but there are some initial studies) being done to analyse written conversations and interactions in order to extract social relationships, in particular power and influence. Furthermore, there is also significant research being done on discovering information from novel data sources, in particular Tweets, and to link the information extracted from text data (e.g. tweets) with more structured data (e.g. about places) available in ontologies from the LinkedOpenData Initiative (e.g. GeoNames ontology). Another topic which is gaining popularity is that of topic detection using a rather old technique known as Latent Dirichlet (LDA). Similarly, research on opinion mining and sentiment analysis, which begun around 10 years ago, is still expanding. The next frontier is to detect fake reviews, and to detect phenomenon such as humour and sarcasm from text. As you imagine, these are relatively complicated tasks since it is sometimes even difficult for humans to recognize sarcastic remarks and to judge whether a review on, say Amazon, is genuine or fake.

Another line of research is that of Question-Answering systems, in which we develop techniques to answer questions expressed in natural language (unlike queries that search engines like Google handle). Question-Answering became popular when the WATSON system, developed by IBM, beat human champions in a TV game show/quiz.

Other communities have also seen the potential in text and data mining, and text and data mining techniques are increasingly being applied. For example, in marketing, a lot of the scientific articles that I read have applied opinion mining (albeit sometimes rudimentary) to find correlations between customer opinions and sales revenue.

There is also tremendous work in text and data mining in Digital Humanities, and I heard of someone in the USA who is studying the evolution of culture based on word trends (i.e. what words were more prominent at a given point in time)

So, yes, in my opinion, text and data mining is being used extensively."

- **How widespread is the use of text and data mining in your own research practices?**

  "For me, my research is in developing algorithms and then applying these algorithms in industrial applications (I work a lot with industry).

  Currently, I am working on opinion mining and also I am using text and data mining to model the linguistic styles on Social Media. For me, I work in this field, and therefore I "use" these techniques a lot. But my usage is somewhat different for someone from marketing for example. That person will be mostly interested in applying text and data mining techniques and extracting the information needed (e.g. opinions), while I would be concerned in improving the accuracy of the techniques and then determining whether higher quality opinions can be mined, and then applying my techniques in an industrial context. Very often, these data from industry demand that the text and data mining techniques be fine-tuned."

- **What is your judgment on text and data mining as a research technique?**

  "Based on my answer to the 1st question, I think that the field is set for expansion. There are number of reasons: Technology has gotten mature many text and data mining techniques are now commercial, robust solutions. More importantly, we are generating a lot of data (big data), and a significant portion of this data is in text format. Some estimates claim that 80% of corporate data is in text format. Therefore, text and data mining techniques are crucial to make sense out of these massive volumes of data.

  Research is constantly pushing the frontier; we now have systems to answer questions, to detect positive and negative opinions in product reviews, to detect social relationships from chat messages, etc.

  All these point in the direction that the field is expected to grow. However, it is quite a pity that academic programs are not doing enough to keep up the pace. More emphasis should be place on 'Analytics' courses, and I am personally introducing a course on Web and Text Analytics and HEC Management School – University of Liege in 2015. There are some schools with such programs already, in the USA, Singapore, and in Groningen (The Netherlands). But these programs focus on the computer science and artificial intelligence (AI) aspects, which in my opinion, discourages students. I had the chance to come from the computer science and AI and also collaborate with industry during my PhD, and this is my reason for pushing for such courses, with a business twist, in the business school where I now work."

- **To what extent do you consider legal protection (copyright) of academic research either helpful or problematic?**

  "In my opinion, academic research should not be copyrighted... but authors should properly cite the works that they use in their works. If someone copyrighted his/her work, then would I be able to replicate his system to evaluate it as a baseline? Furthermore, there is the big question of whether we can patent software? And on what would the copyright apply? On the algorithm? I guess not.

  So, copyrighting should not be done in academic research...if I understood your question clearly. My answer is not applicable for the case if you were asking for 'copyrighting of articles by publishers'."

Prof. Dr. Henrik Boström, professor at department of computer and systems sciences, Stockholm University, Sweden

"One specificity of employments as teacher or researcher at Swedish universities concerns "Lärarundantaget" (see Wikipedia, in Swedish), which makes the teachers/researchers keep the rights to inventions that they have made. You may read more about it here (in English): http://www.managingip.com/Article/1969167/The-professors-privilege.html."

Dr. Maja Bogataj Jančič, LL.M. (Harvard), LL.M. (Torino), Intellectual Property Institute, Ljubljana, Slovenia

"The topic you are dealing with is indeed very interesting and has already become an important part of debates in academic and research communities elsewhere but not much so in Slovenia.

At the moment we are not engaged in any consulting activities with regard to text and data mining. Nevertheless we have identified the increasing text and data mining activity in industrial sector, which is however still at the low rate.

We are currently trying to connect with several Slovenian institutes and organizations, which have more information about text and data mining activity in Slovenia or are engaged in similar processes. In this respect I would ask you to provide me with more specific questions about text and data mining, which would be then sent to them.

As you have already identified intellectual property regime is, of course, a possible obstacle to text and data mining activities in EU. There is no clear exception in the EU or Slovenian copyright law, which would allow researchers to undertake text and data mining without clearing the necessary copyrights of authors of texts and data, which are being mined, as well as other rights, e.g. sui generis database rights of publishers. In Slovenia for example limitations on copyright such as "private reproduction exception" must fulfil certain criteria, the biggest challenge being the general provision enacted in Article 46 of Slovenian Copyright Act, which stipulates that:

"Limitations on copyright are permissible in cases mentioned in this Section, provided that the extent of such exploitation of copyright works is limited by the intended purpose, is compatible with fair practice, does not conflict with normal use of the work, and does not unreasonably prejudice the legitimate interests of the author."

This means that Slovenian courts have a wide discretion for a narrow decision whether certain activity falls within "private reproduction" or some other exception. The same would apply for possible exception with regard to text and data mining.

IP regime as currently in use in EU or Slovenia in this respect may be a serious impediment for a wide use of text and data mining. However no legal actions are seen in practice so far. For this reason further research is necessary to identify the best solution to overcome these problems. Some businesses and as well researchers, who perform text and data mining in Slovenia, have been accepting that text and data mining may not be a legitimate way of using content so they are trying to negotiate a license for their activities or obtain permission.

As mentioned we are trying to connect with institutions and industries, which are directly or indirectly involved in text and data mining and collect a feedback and conduct a research of Slovenian market in this respect."

Published under the editorial responsibility of the Lisbon Council.
The responsible editor is Paul Hofheinz, president, the Lisbon Council.

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

With the support of the European Union: Support for organisations that are active at the European level in the field of European citizenship.

Design by karakas