Dear Commissioners Barnier, Geoghegan-Quinn, Kroes and Vassiliou,

"Licences for Europe –A Stakeholder Dialogue"

Working Group 4: Text and Data Mining

We write to express our serious and deep-felt concerns in regards to Working Group 4 on text and data mining (TDM). Despite the title, it appears the research and technology communities have been presented not with a stakeholder dialogue, but a process with an already predetermined outcome –namely that additional licensing is the only solution to the problems being faced by those wishing to undertake TDM of content to which they already have lawful access. Such an outcome places European researchers and technology companies at a serious disadvantage compared to those located in the United States and Asia.

The potential of TDM technology is enormous. If encouraged, we believe TDM will within a small number of years be an everyday tool used for the discovery of knowledge, and will create significant benefits for industry, citizens and governments. McKinsey Global Institute reported in 2011 [1] that effective use of 'big data' in the US healthcare sector could be worth more than US$300 billion a year, two-thirds of which would be in the form of a reduction in national health care expenditure of about 8%. In Europe, the same report estimated that government expenditure could be reduced by €100 billion a year. TDM has already enabled new medical discoveries through linking existing drugs with new medical applications, and uncovering previously unsuspected linkages between proteins, genes, pathways and diseases [2]. A JISC study on TDM found it could reduce "human reading time" by 80%, and could increase efficiencies in managing both small and big data by 50% [3]. However at present, European researchers and technology companies are mining the web at legal and financial risk, unlike their competitors based in the US, Japan, Israel, Taiwan and South Korea who enjoy a legal limitation and exception for such activities.

Given the life-changing potential of this technology, it is very important that the EU institutions, member state governments, researchers, citizens, publishers and the technology sector are able to discuss freely how Europe can derive the best and most extensive results from TDM technologies. We believe that all parties must agree on a shared priority, with no other preconditions – namely how to create a research environment in Europe with as few barriers as possible, in order to maximise the ability of European research to improve wealth creation and quality of life. Regrettably, the meeting on TDM on 4th February 2013 had not been designed with such a priority in mind. Instead it was made clear that additional relicensing was the only solution under consideration, with all other options deemed to be out of scope. We are of the opinion that this will only raise barriers to the adoption of this technology and make computer-based research in many instances impossible.

We believe that without assurance from the Commission that the following points will be reflected in the proceedings of Working Group 4, there is a strong likelihood that representatives of the European research and technology sectors will not be able to participate in any future meetings:

1. All evidence, opinions and solutions to facilitate the widest adoption of TDM are given equal weighting, and no solution is ruled to be out of scope from the outset;
2. All the proceedings and discussions are documented and are made publicly available;
3. DG Research and Innovation becomes an equal partner in Working Group 4, alongside DGs Connect, Education and Culture, and MARKT – reflecting the importance of the needs of research and the strong overlap with Horizon 2020.

The annex to this letter sets out five important areas (international competitiveness, the value of research to the EU economy, conflict with Horizon 2020, the open web, and the extension of copyright law to cover data and facts) which were raised at the meeting but were effectively dismissed as out of scope. We believe these issues are central to any evidence-based policy formation in this area and must, as outlined above be discussed and documented.

We would be grateful for your response to the issues raised in this letter at the earliest opportunity and have asked susan.reilly@kb.nl (Ligue des Bibliothèques Européennes de Recherche) to act as a coordinator on behalf of the signatories outlined below.

Yours sincerely,

Participants:

Sara Kelly, Executive Director, The Coalition for a Digital Economy

Jonathan Gray, Director of Policy and Ideas, The Open Knowledge Foundation

John McNaught, National Centre for Text Mining, University of Manchester

Aleks Tarkowski, Communia

Klaus-Peter Böttger, President, European Bureau of Library Information and Documentation Associations (EBLIDA)

Paul Ayris, President, The Association of European Research Libraries (LIBER)

Brian Hole, CEO, Ubiquity Press Ltd.

David Hammerstein, Trans-Atlantic Consumer Dialogue

The arguments set out in this letter are also supported by the following individuals and organisations:

Prof Dr Kurt Deketelaere, Secretary General, League of European Research Universities (LERU)

Dr Karl Dittrich, Chair, Association of Universities in the Netherlands (VSNU)

Nicola Dandrige, CEO, Universities UK (UUK)

Prof Kirsten Drotner, Chair, Scientific Committee for the Humanities, Science Europe

Prof Richard Frackowiak, Head Department of Clinical Neurosciences University of Lausanne / Chair, Medicine Sub-Committee of Science Europe

Prof Dr Thomas Risse, Chair of the Academic Committee on Social Sciences, Science Europe

Prof Geoffrey Boulton, Chair of *Science as an open enterprise* report's Working Group, and Chair of the Science Policy Advisory Group, Royal Society

Dirk Inzé, Chair of the Scientific Committee for Life, Environmental and Geo Sciences, Science Europe

Dr. Ir. Véronique Halloin, Secretary General, Fonds de la Recherche Scientifique (FNRS)

Prof Emilio Lora-Tamayo, President, National Spanish Research Council (CSIC)

Professor Rick Rylance, Chair, Research Councils UK (RCUK)

Lisbeth Söderqvist, Associated Professor, Swedish Research Council

Dr Rolf Zettl, Managing Director, Helmholtz Association of German Research Centres

Dr Wim Liebrand, Director, SURF

Prof Martyn Harrow, CEO, Jisc

Prof József Pálinkás, President of the Hungarian Academy of Sciences

Andras Kornai, Computer and Automation Research Institute, Hungarian Academy of Sciences

Professor Sally Wyatt, eHumanities Group, Royal Netherlands Academy of Arts & Sciences

Dr Wilhelm Krull, Secretary General, Volkswagen Foundation

Prof Dr Jos Engelen, President of The Netherlands Organisation for Scientific Research (NWO)

Dr Petr Škyřík, National Cluster of Information Education, Czech Republic (NAKLIV)

Sir John Sulston, Institute for Science, Ethics and Innovation (iSEI)

Dr Simon Chaplin, Wellcome Trust

Dr Tim Hubbard, Wellcome Trust Sanger Institute

Dr Rolf Apweiler, Associate Director, European Bioinformatics Institute (EMBL-EBI)

Alicia López Medina ,Confederation of Open Access Repositories

Alma Swan, Director of Advocacy Programmes, SPARC Europe

Dr Victor Henning, Co-founder & CEO, Mendeley

Dr Harald Müller, Library Director, Max Planck Institute for Comparative Public Law and International Law

Associate Professor Dr. Lucie Guibault, Institute for Information Law (IViR), University of Amsterdam

Prof Dr Mireille van Eechoud, Institute for Information Law (IViR), University of Amsterdam

Dr Stef van Gompel, Institute for Information Law (IViR), University of Amsterdam

Prof Charlotte Waelde, Chair in Intellectual Property Law, University of Exeter

Prof Lilian Edwards, E-Governance,University of Strathclyde

Prof Dr Rainer Kuhlen, Chair, European Network for Copyright in support of Education and Science

Teresa Hackett, Electronic Information for Libraries

David C Prosser PhD, Executive Director, Research Libraries UK

Janet Peters, Chair, Research Libraries UK

John Dolan, Chair of Council, Chartered Institute of Library & Information Professionals

Tim Padfield, Chair, Libraries and Archives Copyright Alliance

Kimmo Tuominen, President, Finnish Research Library Association

Kristiina Kontiainen, Finnish Library Association

Ap de Vries, CEO, Netherlands Public Library Association

Bas Savenije, President, FOBID Netherlands Library Forum

Andrew Green, CEO, LlyfrgellGenedlaetholCymru / National Library of Wales

Joy Palmer, Library and Archival Services, Mimas, University of Manchester

Ann Rossiter, Executive Director, Society of College, National and University Libraries

Michel G Wesseling, President, NVB: The Dutch Association of Information Professionals

Natalia Manola, University of Athens, Greece on behalf of OpenAIRE

Dr. Evi Sachini, Head, Strategic Planning & Development Department, National Documentation Centre Greece (EKT)

Kevin Ashley, Director, Digital Curation Centre (DCC)

Dr. Lukasz Bolikowski, Centre for Open Science, ICM, University of Warsaw

Simon Tanner, Deputy Head of Department of Digital Humanities, King's College London

Christoph Kratky, President, FWF - Der Wissenschaftsfonds, Austrian Science Fund

Maël Brunet, Head of Office, OpenForum Europe

Estelle Derclaye, Professor of Intellectual Property law, University of Nottingham

Wojtek Sylwestrzak, Centre for Open Science, ICM, University of Warsaw Jakub Szprot, Centre for Open Science, ICM, University of Warsaw

Daniel Dietrich, Chair, Open Knowledge Foundation Deutschland

Jacek Maj, Director, Collegium Artium

Annex

1. International Competitiveness

A healthy diversity of SMEs and a solid research base that makes the most of new technologies is a must for a vibrant European economy. Given the EU's comparatively low levels of R&D investment there is an urgent need to remove barriers to growth, particularly we would argue in the technology and research sectors. There is a significant body of evidence that shows that the adoption of the potentially life-changing technology of text and data mining is being severely hampered by market failure, a lack of legal certainty, and the division of information into silos [1]. These barriers, however, do not exist in the United States, Japan, Israel, Taiwan and South Korea [2] – countries that already have much higher levels of R&D intensity than the EU [3]. Due to limitations and exceptions in copyright law it is permissible for computers and servers based in those countries to freely read any copyrighted works they have legal access to with no requirement to seek out further permissions, negotiate and potentially pay for extra licences. Ironically, given the international nature of copyright, this access also includes copyrighted works from Europe, which we ourselves cannot currently freely use in the same manner due to the lack of copyright flexibilities in EU law. We think it vital that the researchers and companies of Europe are able to compete on a level playing field with our global competitors, and not suffer further impediments to reusing knowledge that we already have legal access to.

1. The Value of Research to the EU Economy – The Need to Prioritise

We value highly the important role that publishers play in the creation and dissemination of scholarly information which in 2008 was estimated [4] to represent €7.68 billion worth of investment globally. In comparison the global investment by researchers in the undertaking and communication of the same scholarly information was €210 billion in the same year, and the EU's public investment in scientific, medical and other forms of research according to OECD totalled € 90.7 billion [5].The economist Professor Jonathan Haskel at Imperial

College London has provided empirical evidence of how investment in scientific research contributes to the long term economic growth of a country. [6] In the context of the debate on barriers to the adoption of TDM, it is also interesting to consider the well proven link between the adoption of ICT technology and economic growth [7].

Even in purely economic terms, aside from the vast social and health benefits that result from research, it is clear that supporting the capacity of research to use all opportunities available to it must be the overriding priority in any discussion of those technologies. It is essential that the EU is not unnecessarily restrictive in its regulation of these technological tools.

Furthermore, the EU must also be cognisant of the fact that by essentially trying to create a new restricted act in law for TDM it seeks to control whether a researcher or technologist's computer can without further permission analyse and read material it already has legal access to. This raises not only serious economic and legal questions but ethical ones also, as it places in the hands of the licensor power over what computer based medical, scientific and other forms of research can and cannot be undertaken downstream. This raises issues of the utmost seriousness for the research sector.

In the global marketplace for research, the more difficult it is to use computer-based analytical tools in the EU, the more likely it is that the EU could lose research, research funding, and our current international position which may well not be recoverable. Similarly European technology companies will be at a significant disadvantage compared to other countries and trading blocs around the world, thus impacting on growth.

1. Impact on Horizon 2020

H2020 establishes a single strategic framework for research and innovation, recognising that an integrated approach will help to work towards solving societal challenges and maximise the competitive impact of research and innovation. It is within this context that H2020 will seek to fund new technologies, innovative research in SMEs, and spreading excellence in science. TDM has the potential to impact all of these areas and its development could play a key role in addressing societal challenges in the future. The H2020 programme will be finalised over the course of 2013. The discussions that are occurring in parallel under the Licences for Europe initiative have the potential to impact directly on the effectiveness of H2020. At a point when one arm of the Commission is making efforts to spread excellence in science by setting rules for open access to research publications and data, as well as planning for the achievement of an 'online' European Research Area, other Commission initiatives should not counter this effort by establishing further barriers to research and innovation. It is therefore essential that DG Research is included as part of Working Group 4.

1. The Open Web and Big Data

In an environment of legal uncertainty, some EU technology companies and research organisations already risk legal and financial sanction and mine the open web – the largest single database the world has ever known. We believe this activity must be lawful in order put us on an equal footing with our global competitors. In the context of Licences For Europe,we are extremely unclear who - other than perhaps the government as was concluded in Japan – is in a position to "grant permission" for the mining of the open web and would be most grateful for clarification on this crucial point.

We would also like to point out that we were somewhat surprised that only one organisation, COADEC, represented European technology companies at the meeting. The rest of the technology companies present were US in origin.

1. The Extension of Copyright Law to Analysis and Reading

We strongly support and recognise the importance of the time-limited monopolies that copyright law creates. We also recognise that non-Open Access publishers have concerns about the unauthorised distribution of their works, and that publishing is an essential function in the research cycle.

Copyright law is an exclusive right intended to allow the creator of a work to develop and exploit the marketplace for his or her work, as he or she sees fit, by preventing uses which conflict with the normal exploitation of those works. This rationale however does not readily apply when the onward usage of that work is in no way substitutable and does not rival the commercial viability of the original work. We are not aware of any evidence to indicate that through extending the restrictions of copyright over TDM technologies and limiting the use of technological tools to read research, that it will result in larger markets, more research being published in the long term, or produce any other net public benefits. The opposite result appears much more probable.

Copyright law was not developed to restrict how works are read, or the processes that humans undertake to develop new thoughts and ideas [8]. As pointed out above TDM does not trade on, or replicate the underlying expressive purpose of the copyright work, and if performed using a pen and pencil is totally unregulated by copyright law. To quote "Digital Opportunity" [9], which the UK government has built its current copyright policy in this area on, "*the technology provides a substitute for someone reading all the documents. This is not about overriding the aim of copyright – these uses do not compete with the normal exploitation of the work itself –indeed, they may facilitate it. Nor is copyright intended to restrict use of facts. That these new uses happen to fall within the scope of copyright regulation is essentially a side effect of how copyright has been defined, rather than being directly relevant to what copyright is supposed to protect.*"***[10]***

The law is clear that it does not seek to regulate the building blocks of knowledge – facts and data. The Database Directive (96/9/EC) itself, repeating Berne, TRIPS and other international legal instruments states that "*the right to prevent unauthorized extraction and / or re-utilisation does not in any way constitute an extension of copyright protection to mere facts or data*" [11]. Given that the product and output of TDM is the extraction of facts, organisation of documents, or simply a number or word indicating a finding or correlation between variables [12], we simply do not believe that copyright law or the database rightare relevant here and in the context of Working Group 4 call for an evaluation of whether copyright and the database right can or should be extended to cover the "reading" of databases and the utilisation of facts.

In short we wonder how such outputs or hypotheses can be substitutable or conflict with the normal exploitation of the text, and how it conflicts with the legitimate interests of their original author when legal access has already been given.

[1] Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute. 2011. http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation

[2] Text Mining and Data Analytics in Call for Evidence Responses. UK Government http://www.ipo.gov.uk/ipreview-doc-t.pdf

[3] Value and Benefits of Text Mining. Dr Diane McDonald. Joint Information Systems Committee. 2012. http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

[1] Value and Benefits of Text Mining. Dr Diane McDonald. Joint Information Systems Committee. 2012. http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

[2] The US, Israel, South Korea and Taiwan assert the Fair Use doctrine for this activity. Whereas Japan introduced in 2009 alongside other limitations and exception aimed at boosting their internet economy a specifically designed limitation and exception to permit TDM.

[3] Innovation Union Competitiveness Report 2011 – Investment and Performance in R&D. European Commission. http://ec.europa.eu/research/innovation-union/pdf/competitiveness-report/2011/part_1.pdf

[4] Activities, costs and funding flows in the scholarly communications system in the UK Report commissioned by the Research Information Network (RIN). 2008. www.rin.ac.uk/system/files/.../Activites-costs-flows-report.pdf

[5] Government budget appropriations or outlays for R&D for 24 of27 EU member states 2008. Data extracted on 11 Feb 2013 15:36 UTC (GMT) from OECD.Stat

[6] How much does publicly funded research contribute to UK economic growth? J Haskel.27/01/2010.

http://www.ceriba.org.uk/pub/CERIBA/CeribaPublicfundedresearch/Haskel_publicly_funded_research_and_economic_growth.pdf

[7] M Franklin, P Stam and T Clayton.ICT impact assessment by linking data across sources and countries. Office of National Statistics.

[8] When Copyright Law and Science Collide:Empowering Digitally IntegratedResearch Methods on a Global Scale. Jerome H Reichman& Ruth L Okediji.Copyright and Science. 2012. http://scholarship.law.duke.edu/cgi/viewcontent.cgi?article=5351&context=faculty_scholarship

[9] http://www.ipo.gov.uk/ipreview-finalreport.pdf

[10] Echoing the Hargreaves Review the Japanese Government before introducing a limitation and exception in 2010 for TDM stated "*In an advanced information society amidst*

vast volumes of information, data analytics technology, which allows the extraction of information as well as the advanced processing of such knowledge, is a necessity for users, as well as a fundamental of a digitally networked society. It can also be argued that the development of research involving data analytics has many societal benefits. In addition, another side to the argument is that research developments using data analytics do not use the (artistic) expression contained in a copyright work itself, as it is no more than the extraction of information. And that while in the process of data analytics a copyright work is used, its actual essence is not."

[11] Please see the following from international treaties and how copyright law relates to "expression" rather than "ideas" and "facts" - Berne Art 2(8), TRIPs Art 9, TRIPs Art 10, WIPO Copyright Treaty Art 2 and Copyright Treaty Art 5. In addition we note that Stockholm Revision Conference states "*news items or the facts themselves are unprotected*" and the WIPO Guide (1978) states: "*The rationale of this provision is that the Convention does not set out to protect mere news or miscellaneous facts because such material does not possess the qualifications necessary for it to be considered a work.*"

[12] Please see attachment which shows an example output from text and data mining. In this case it shows a previously unknown relationship between the protein e cadherin and Parkinson's disease. This strong correlation shown by a difference of over 5 (8.3820 – 13.89) and is derived using statistics from what has not been written about in the text.