Openforum Academy

ROUND TABLE REPORT

HOW TO UNLEASH THE INNOVATIVE POTENTIAL OF TEXT AND DATA MINING IN THE EU

22nd of October 2015, Brussels

Report

ROUND TABLE: How to unleash the innovative potential of text and data mining in the EU

Brussels, 22nd of October 2015

Hotel Silken Berlaymont, Boulevard Charlemagne 11, B-1000 Brussels, Belgium

DISCLAIMER

This report was prepared by our rapporteur, Diana Cocoru, for OpenForum Academy (OFA). The summaries of the speakers' introductory presentations and of the ensuing discussions presented in this report are based on the rapporteur's notes. Although every effort has been made in good faith to reflect and to convey objectively the essence of the speakers' individual presentations and the associated ensuing discussion, the summaries set forth below are not in any way binding or necessarily complete. Moreover, the views expressed in this report do not necessarily reflect those of the rapporteur or of OFA. Neither the rapporteur nor OFA should be held accountable for any claimed deviation from the original speeches.

OpenForum Academy gratefully acknowledges IBM's and Oracle's sponsoring of this event. OFA welcomes financial support for its events, but always maintains independence of the discussion itself and the follow-up White Paper.

SPEAKERS

Catherine Stihler | Member of the European Parliament, IMCO Vice-Chair and substitute in ECON. Rapporteur of the IMCO Opinion for the EP's own initiative report on the implementation of the InfoSoc Directive

Peter Murray-Rust | Founder of ContentMine and Shuttleworth Fellow, member of the advisoryboard for the Open Knowledge Foundation

Jean-Francois Dechamp | Policy Officer at the European Commission, in DG Research and Innovation

Complete recordings of the various speakers' introductory speeches are available online, on OpenForum Europe's Youtube channel.

MODERATOR | Karel De Vriendt, Director, OpenForum Europe

RAPPORTEUR | Diana Cocoru, Senior Policy Analyst, OpenForum Europe

Other details of the event, and the speakers' presentations, are available here.

CREDITS

Photo and White Paper "How to unleash the innovative potential of text and data mining in the EU" are attributed to OpenForum Europe, under license CC BY SA 4.0.

EXECUTIVE SUMMARY

The reform of the Copyright Directive 2001/29/EC has been one of the focal points for the previous and current Commissions, as well as the current European parliamentary term. Several key documents, such as the European Parliament's own initiative report on the implementation of the Copyright Directive or the Commission's Digital Single Market Strategy, shed some light on where the European institutions position themselves in relation to text and data mining ('TDM'). Ongoing projects (such as 'FutureTDM', an EU-funded Horizon 2020 project which started in September 2015), are examples of how the Commission seeks to obtain feedback and knowledge from the ground.

In order to grasp the difficulties of enabling the right legal environment for TDM, one of the speakers on the panel presented in very practical terms how researchers proceed in order to extract information from thousands of papers, and associated obstacles which they can face.

The debate also covered the main stumbling blocks when it comes to unleashing the innovative potential of TDM activities. The participants engaged in discussions about publishers' resistance to change their business models, their fear of losing control over the content, the negative impact of imposing the use of certain mandatory mining APIs, as well as the existing legal uncertainty around what is allowed and under which conditions.

This paper looks at the current and future EU legal and policy frameworks governing TDM. It then presents how researchers use this mining activity to discover and then to scrape through thousands of papers, and concludes with a section focusing on certain obstacles which can block the innovative potential of TDM.

HOW DO EUROPEAN INSTITUTIONS POSITION THEMSELVES IN RELATION TO TEXT AND DATA MINING?

The reform of the Copyright Directive 2001/29/EC is one of the focal points of the current European parliamentary term. The issue of TDM is extremely complex, with several associated misconceptions. However, thanks to institutions such as the European Parliament, and also to work initiated by the Commission (e.g., through 'Licenses for Europe') a couple of years ago, it seems that TDM has become a little more of a focus item and received more attention. It still remains the case that TDM represents only a small aspect in the realm of copyright, even if the Commission, the European Parliament, and the Member States themselves agree that research and innovation are important.

It is widely understood that research and innovation are the drivers of European competitiveness and that TDM is a useful tool, which speeds up data processing and analysis and helps study data in literally any sector, ranging from agriculture to chemistry as well as all other sciences, the arts, humanities, and social disciplines.

Ms Stihler was the rapporteur of the IMCO Committee for the Opinion on the European Parliament's (EP) own initiative report on the implementation of the Copyright Directive, adopted in July 2015, and also a permanent member of the copyright Working Group set up by the Committee of Legal Affairs in the European Parliament. She underlined that during the policy process, there were varying views across the political spectrum and nationalities on a whole group of areas. In the final resolution, the Parliament stressed "the need to properly assess the enablement of automated analytical techniques for text and data mining for research purposes, provided that permission to read the work has been acquired." Ms Stihler underlined that despite the advantages which TDM can bring, legal uncertainties around its use, as well as practical difficulty of obtaining access and permission to mine, are seen as factors slowing down the process of data handling. This is why some argued that the wording should have been stronger, calling for a mandatory exception permitting TDM across the EU.

Along the same lines, in October this year, the European Parliament Research Service has published a study on the review of the copyright framework, acknowledging the lack of certainty over the TDM exception, and pointing out that this represents a potentially serious gap in the EU *acquis*.

Mr Dechamp took the opportunity to explain what the European Commission is doing in regards to TDM. Part of the Commission's job is to fund researchers and projects in different fields. When it comes to research and innovation, the aim is to optimise the impact of publicly funded research. This is why TDM is seen as an optimisation of research, because automated techniques reduce the time needed to analyse large sets of data. The EU-funded 'FutureTDM' Horizon 2020 project [1], which started in September 2015, is one example of how the Commission tries to obtain feedback and knowledge from the ground.

Mr Dechamp underlined the importance of acknowledging that there is a big difference between the life of researchers and that of producers of cultural products: whilst scientific researchers are both the producers and authors of goods, they are also consumers and readers. This is not necessarily the case for those who consume cultural products, for instance. As authors, scientific researchers typically are not paid, and do not receive any kind of compensation or royalty (over and above the remuneration which they receive from universities to do their job, which includes the dissemination of results to their peers). This brings a different aspect to the issue of copyright.

The European Commission has different strategies, one of which being the Digital Single Market, published on the 6th of May this year. The section on TDM talks about innovation and research for both commercial and non-commercial purposes. The Commission also underlines the unclear legal framework, the divergent approaches at the national level, the need to enable researchers to use a wider range of materials, and the need for cross border collaboration, knowing that research communities are generally international.

Turning to the Council, the conclusions published in May 2015 made a declaration on open and data intensive research, talking about the promotion of innovation driven by TDM, taking into account research needs.

All the above shows at least a basic consensus among the three main European institutions that TDM is a topic which needs to be addressed. The Commission wants to be pragmatic, it wants to target modern copyright legislation. What is currently being proposed is to have a Communication by the end of this year that will set the scene and will put forward the working plan for the modernisation of the EU copyright environment. That would be the first step to tackling a few of the specific copyright issues, which are not all of direct relevance to TDM.

1. Link to the FutureTDM project web site: http://project.futuretdm.eu/project-overview/

However, the second step, which is scheduled for spring 2016, is to have a complete set of legislative proposals and solutions. That is where TDM will be tackled in specific terms. The objective is to give European researchers and innovators the best conditions in which to do their jobs, argued Mr Dechamp.

WHAT DOES TDM LOOK LIKE IN PRACTICE?

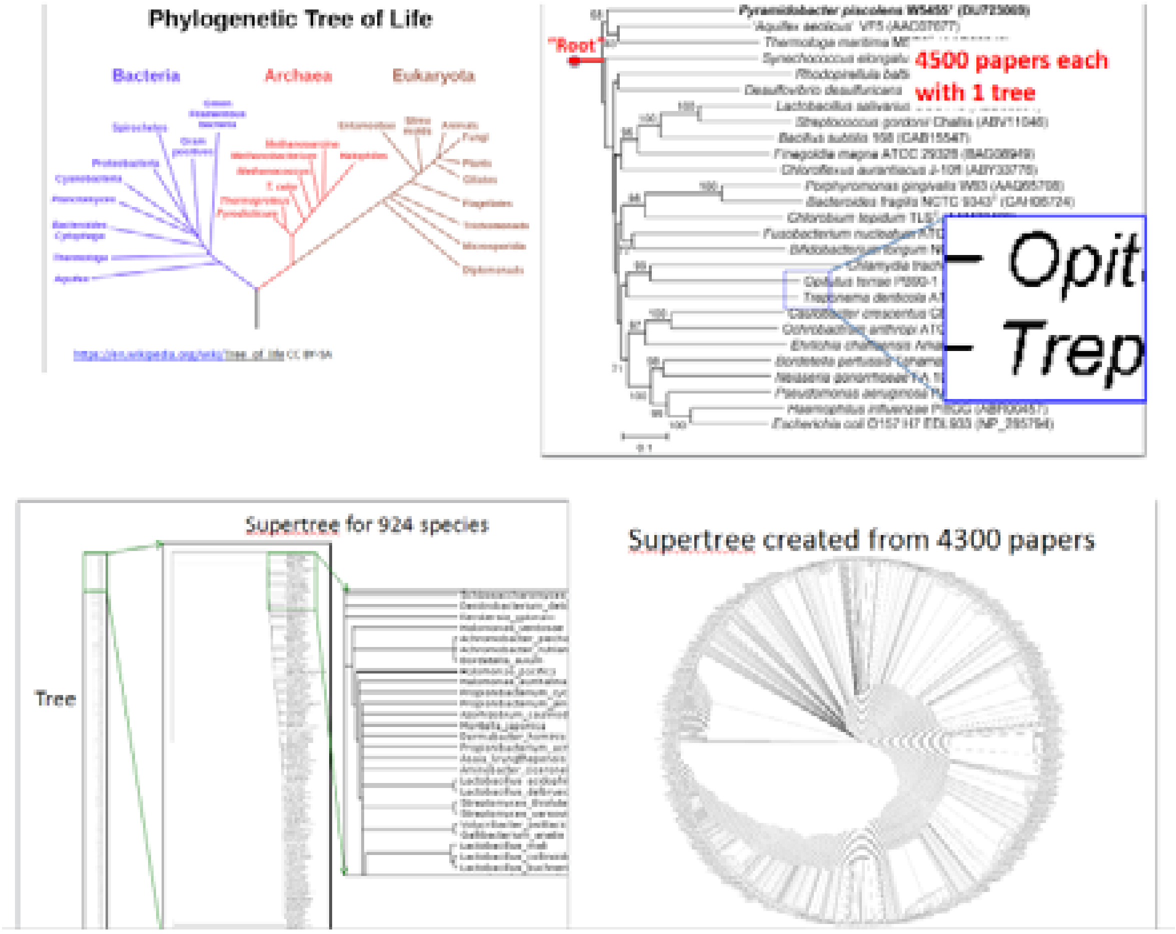
The first function of TDM is that of discovery. Peter Murray-Rust took the example of the MRC Epidemiology Unit at the University of Cambridge, which is interested in (inter alia) all literature relating to obesity. The researchers from this Unit have to read 10,000 papers and decide which of them are valuable. They have 20 seconds to read each one and to decide whether it is about obesity or not. The language used in these papers includes terms such as 'body mass', 'mortality risk', and so on. Reading 1600 papers in 2 and a half days, together in a team of six people is much less efficient than a machine doing it. Mr Murray-Rust founded the Content Mine project, which consists entirely of open software and allows this kind of work to be done. The software is made available under the terms and conditions of the Apache License version 2 or the BSD equivalent and it can do this reading of 10,000 papers in a few minutes.

Mr Murray-Rust gave a second example, to do with EU clinical trials, of which some 400,000 are published in government repositories. Activists in the UK have been pushing for trials to be open. The problem with not publishing them is that if one only publishes the positive trials about a drug and not the negative ones, then public health will suffer. This is how it works in practice: every clinical trial is allocated its own unique number, which gets published in the EU registry. However, there is nothing in the registry to indicate what the follow-up (if any) has been. In this case, the literature can be read to search for each unique number (i.e., the index of each clinical trial) in order to find out about the follow-up of those clinical trials. This is made possible by TDM.

The second function of TDM is to scrape the identified papers. Mr Murray-Rust described how they search and analyse papers, and how they carry out complex transformations on them. This is what is known in the US as 'transformative use', which is allowed under the 'fair use' provisions of the US copyright law regime. In the US, if one takes a document and transforms it by adding value' to it, the act of creating the transformation is not prohibited by copyright.

Mr Murray-Rust asserts that the TDM activities at Cambridge which he described should fall into the same category. According to him, 'the right to read is the right to mine' and no further obstacles should be imposed. When speaking about TDM, Mr Murray-Rust prefers to speak about 'content' mining. This is because although some text will typically be included in any scientific paper, most of the rest of the paper will consist of diagrams, mathematical equations, facts and tables. That is often the most valuable part of the paper for a researcher, according to Mr Murray-Rust, and he therefore uses the term "content" as opposed to 'text and data', because he perceives a risk that otherwise publishers could claim that because something is a creative graphic work (rather than just text), it belongs to them. Instead, Mr Murray-Rustasserts that in the case of a factual representation, the mere presentation of a fact in a diagram does not mean that the presentation of that fact becomes a copyrighted piece of creative work; it still remains a simple fact, which belongs to the public domain.

Mr Murray-Rust took the example of the E. coli bacterium, to show what researchers can do by aggregating information. In this example, he built a tree of this bacterium, using 5,000 papers. The team of researchers read the literature, extract the relevant images out of each paper, conduct some very advanced analytics and transform the extracted data into something which resembles the images below. The process goes through several stages, and ultimately the end result is a tree of bacteria that has been published in the literature. To do this, Mr Murray-Rust says that the copyright laws have been violated. Researchers cannot do reproducible mining without violating copyright. They cannot present their findings, because if they were to publish them, they would need to reproduce and publish material in which copyright has been claimed. This is because even when the act of data mining (as in the UK, with the recent exception) is allowed, that permission only allows publication of the results (with accreditation where feasible). What is not allowed is a public reproduction of anything other than the facts which mining reveals or extracts (knowing that facts are not protected by copyright). Therefore, to the extent that publication or reproduction of the mining process itself would expose or reveal or copy anything other than unprotected data, then such publication or reproduction of the mining is likely to infringe copyright of such richer material. That is the real problem which scientists face: to be prevented from publishing what they feel they need to publish.


Phylogenetic Tree of Life
Bacteria
Archaea
Eukaryota
"Root"
4500 papers each
with 1 tree
Opit
Trep
Supertree for 924 species
Tree
Supertree created from 4300 papers

Source: Peter Murray-Rust's presentation available on SlideShare.

WHAT BLOCKS THE INNOVATIVE POTENTIAL OF TDM FROM BEING UNLEASHED?

Legal uncertainty

The situation of legal uncertainty has partly been addressed in the UK, by the recent inclusion of an explicit copyright law exception (which cannot be overridden or waived by contract) for content mining, albeit with some limitations. As of today, the UK is the only EU Member State to have enacted such a specific exception. As explained by Ms Stihler, the UK exception allows researchers with lawful access to a copyrighted work to make copies of the work for the purpose of text and data mining, but only for non-commercial research. It requires attribution of the source, and prohibits the sharing or selling of the copies. Looking at the impact that TDM has on research, Ms Stihler pointed out that in Scotland, European funds of approximately 530 million EUR (from the previous framework programme, FP7) went directly towards research and technological development.

Ms Stihler wanted to investigate further about what researchers have to say about this. Therefore she wrote to all the higher education establishments in Scotland (19 in total), of which 16 are more focused on research. She asked them to express their views on whether they consider that the Copyright Directive should be reviewed, and to share their thoughts on the TDM exception. Currently, only around half of them have replied; all of the responses received welcomed upcoming reform on copyright at the EU level, as well as welcoming the recent UK-specific exemption of TDM for non-commercial purposes. Several respondents pointed out that they would support the inclusion in the new copyright rules of a specific exception for TDM for non-commercial purposes, provided that this did not weaken today's UK-specific exception. Ms Stihler noted that views were divided as to whether to extend the TDM exception to commercial purposes, with some universities being strongly in favour, and others showing more caution. She highlighted that the responses received to date how that it is often difficult in a university context to distinguish between 'commercial' and 'non-commercial'. Researchers need a clear definition of what constitutes TDM for non-commercial purposes and for commercial purposes in order for them to be able to distinguish between the two in practice.

Often universities work together with industries and pharmaceutical companies. Frequently, medical innovation starts from a university's analysis, and is then implemented by commercial organisations. Therefore, the overall exchange of knowledge between universities and the public, private, and third sectors is continuously increasing. Ms Stihler reported that this growth rate in cash terms in the UK is around 5% p.a. [2], and that over the longer term, income has risen by about 45% since 2003. This clearly shows that knowledge transfer from universities to industry facilitates scientific, environmental, and health innovation.

The distinction between commercial and non-commercial purposes, in the UK exception for TDM, arises from two places. First of all, the distinction appears in the provisions of European laws, and also in national laws from some Member States, such as the UK. Adopting a TDM exception limited to non-commercial uses in the UK was justified by the fact that such a change could be done through a statutory instrument instead of passing through the UK Parliament's classic procedure (since the provision already existed in the EU copyright legislation), and thus increase the chances of rapid adoption. The second place it comes from, is the Creative Commons, which does not deal only with science, but also with music, culture etc. And in those fields, there is a valid distinction between commercial and non-commercial. The problem is that nobody can define non-commercial, and several papers have been written, which said very clearly that 'non-commercial' cannot effectively be defined in a scientific context, and therefore should be removed if possible. The difficulty to distinguish between commercial and non-commercial is also reflected in a recent case in Germany, whereby one lower court found that teaching was a commercial activity, and the higher court overturned the decision and ruled that teaching is a non-commercial activity.

The main conclusion along this line was that adopting a modern copyright reform fit for the digital world, with certain mandatory exceptions that embrace innovation, should be at the forefront of common thoughts and actions. And especially so, in the light of universities facing 20%, sometimes even 40%, cuts in their budgets. There are ongoing efforts to enable good research to be undertaken, but the issue needs to become more mainstream and MEPs should come out and meet researchers, since so much could be gained by improving communications.

2. According to the Higher Education Council for England

Publishers' resistance

The audience agreed that there were different types of publishers and that interaction with them also varies. There are the federations, such as the International Association of STM publishers (STM-Publishing) or a number of associations representing open access publishers (OASPA, the Open Access Scholarly Publishers Association). Some of the major publishers are also individually talking to European legislators. Therefore there was a consensus in the audience that not all publishers are the same – thus generalising about "publishers" is neither appropriate nor accurate.

Traditional publishers tend to put licence-based solutions on the table. Academics and research librarians, who are negotiating directly with publishers in order to obtain access to the research, generally feel that licences are not the solution, and that they are not best adapted to the 21st century.

DG Research had contracted a group of experts to analyse standardisation in the area of innovation and technological development, notably in the field of TDM. Among other conclusions, these experts found that licences are suitable, but only in the short term, until a better solution is found. Getting back to the main issue, which is the copyright of the authors, it was pointed out that often authors/researchers have to give their copyright to the publisher, in which case the publishers naturally become the right holder.

One of the drawbacks of the licences which publishers are pushing for is that they entail additional costs: time and money for researchers to request the licences and to handle them, a process which in the end can prove unsuccessful. Another aspect is the inter-disciplinarity of some research, which can require different licence versions or variants because the content to mine comes from different sources. One may have some specific licences with one publisher, which might not be compatible with another publisher's licences. Researchers want to be able to combine data and text publications from different right-holders, which means that the negotiation of all the required licence terms and conditions is likely to prove to be a very long process. There was a consensus among the participants that publishers often take a very long time to respond to individual licence enquiries; in addition, there are hundreds of different scholarly publishers.

Some participants put forward the idea that one of the 'real reasons' for publishers' resistance is their feeling that publishers are an industry 'under threat'. The problem is that disseminating content is becoming virtually free of cost, and publishers are selling content at a huge cost, thus the world will eventually question what value is added here by publishers. The second, and more insidious, reason is that some publishers have downstream, secondary products for added value, which means that any provider of secondary databases represents a significant threat for them.

Another reason for the publishers' resistance, besides the fear of losing income that otherwise would be generated by TDM-specific licences, is fear of loss of control over the content which is being downloaded for mining purposes.

Academics and researchers are trapped in this power struggle by the prestige factor: having published in a famous journal seems to still be more important to most researchers than simply releasing the paper through a less prestigious open access journal; and also for prestige-based reasons. Universities themselves often ask researchers to publish in a "respected" journal.

A solution to all these problems was suggested during the discussions: to urge all public organisations to maximise the scientific value of publications by releasing everything fully on CC BY or CC0 terms (which attempt - as far as possible - to dedicate the work to the public domain; however, in some jurisdictions, one cannot divest oneself of one's ownership). The Commission has already put some rules for open access in place, in the Horizon 2020 programme, and is now encouraging researchers to keep their copyright and grant a licence for publication to their publisher, mentioning good example being CC BY licences. For the mid-term review of the Horizon 2020 programme, planned for 2017, the Commission intends to strengthen its requirements for openness in general.

When the question of how the suggestion to keep copyright is received by the community was raised, it was pointed out that some individual authors with forceful personalities have reportedly refused to assign their copyright and have succeeded in securing their publisher's agreement. Some publishers universally refuse, but ultimately researchers have to ask themselves what is the purpose of the publication: to advance a scientist's career or actually to help save peoples' lives and/or do better science? In practice, 85% of medical funding is wasted because it is not published at all, or it is not published properly, it is duplicated or the design is bad.

Closed publication means bad science, and although many young people realise this, they do not always have the capacity, the bargaining power, the authority or the perseverance to act on it.

There seemed to be an agreement amongst those present that one group which has the potential to change today's equilibrium is the funding community, e.g., by insisting that no funding or grant will be advanced for a research project unless each researcher keeps his or her copyright.

The negative impact of imposed APIs

The last two years have shown that in various ways publishers are actively increasing their efforts to control this market: by licences (as seen above); or by imposing APIs; or by imposing the use of specific portals to access their information. One problem with imposed APIs is that this practice allows publishers to control what researchers do. Consequently, researchers do not have certainty when they use that interface that they obtain access to the totality of the paper which they could have accessed through the website. Another perceived risk is that imposed APIs could become a perfect gateway for monitoring what researchers are doing.

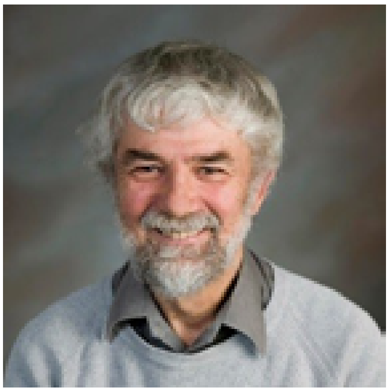
Mr Murray-Rust highlighted that some publishers are building monopolies on the infrastructure, and they will end up with something similar to Apple's universal infrastructure if APIs imposed by publishers become the norm. Further, he gave a typical example of how APIs can limit the researcher: through the introduction of a "captcha" validation routine. For example, after every 25 papers downloaded, the user has to stop and type in the captcha, considering that publishers usually set limits of e.g. 100 papers a day. Thinking of the case study above, with the clinical trial and the 100,000 papers, this would take three years to go through.

One rationale that some publishers put forward to justify their request to use their APIs is the fact that not using them would break their services. However, in order to use their APIs, the user is required to accept the terms and conditions of a contract, which may purport to limit the access of the researcher to only parts of the needed papers, or which purport to limit the re-use of the outcome of the TDM research activities. Even worse, publishers are now getting to the stage where they are uniting and producing their own APIs to "help" researchers. The problem with publisher's APIs is that they are not designed with the purpose of giving an open, transparent access, but a closed and limited one. They may be more efficient, but that is a big price to pay for liberty.

SPEAKERS' BIOGRAPHIES

Catherine Stihler, *MEP from Scotland*

MEP since 1999, Ms Stihler is IMCO ViceChair and substitute in ECON. Ms. Stihler is a strong advocate of TDM. She was the rapporteur of the IMCO Opinion for the EP's own initiative report on the implementation of the InfoSoc Directive. When the report was adopted in July 2015, she underlined that the "impact of TDM on research is phenomenal" and that she hopes to see a new exemption for TDM at an EU level and an assessment made concerning TDM for commercial purposes.



Peter Murray-Rust, *Founder of ContentMine and Shuttleworth Fellow*

Mr. MurrayRust is a renown chemist, as well as longtime advocate for open access and open data. His research has been focused on the automated analysis of data in scientific publications, as well as the formation of virtual communities like the Virtual School of Natural Sciences in the Globewide Network Academy. As part of his campaigns for open data in the sciences he is on a the advisory board for the Open Knowledge Foundation. He is also a cofounder of the Blue Obelisk movement, which advocates for open access, open source, and open data. Currently he is Chemist at the University of Cambridge.

Jean-Francois Dechamp, *Policy Officer at the European Commission*

Since 2005, Jean-François Dechamp has been Policy Officer in the Directorate-General for Research and Innovation of the European Commission in Brussels, Belgium. He has helped shaping the policy related to open access and, and more recently, to copyright in the remit of Text and Data Mining activities (TDM). He has worked previously for the pharmaceutical industry and for various non-governmental organizations in the field of health and humanitarian aid. He is a Doctor of Pharmacy from the University of Strasbourg, France.

Legal Information

OFE Limited, a private company with liability limited by guarantee
Registered in England and Wales with number 05493935
Registered office: 1 Blunt Road, South Croydon, Surrey CR2 7PA, UK