Annex 1 – Full case studies

CASE STUDY 1: CONTINUING UNCERTAINTY

Aggregating the world's open access research papers

CORE (www.core.ac.uk) is a UK initiative which aims to aggregate all open access research outputs from repositories and journals worldwide and make them available to the public. With over 36 million metadata records, and 4 million full-text documents, CORE is now one of the largest databases of scientific literature in the world.

An uncertain start

Launched in 2011 by the Open University, with support from Jisc, the project's early years were dogged by legal wrangling over copyright. As its founder, Petr Knoth, explains, 'The massive difference the exception has made for us is that in the past we would repeatedly be asked "Can you actually do this?" and "How do you know that the articles are open access?"'. As there are no robust standards to indicate the license applicable to particular articles, proving that CORE had the legal right to take copies of millions of articles was no easy task.

Thus, significant management time had to be devoted to resolving legal disputes and formal complaints, including from other academic institutions within Europe. 'One particularly difficult case probably took two months of my life to resolve,' says Knoth, 'and all for no benefit – but with the exception in place, that no longer happens'.

A bright future

The UK's exception, introduced in 2014, has removed the risk that material is copied inappropriately. In turn, this has freed up Knoth and his team to focus on growing CORE's coverage: 'I no longer need to go to many meetings to discuss whether we can do this or that – that's the benefit for me,' he says.

Constraints on commercial potential

Innovation Engineering, an Italian small and medium-sized enterprise (SME), now uses CORE to power Wheesbee (www.wheesbee.eu) - a one-stop-shop to search, retrieve, organise, and share web-based technology related information. Another user is ResearchResearch, a commercial provider of academic intelligence services. Its CEO, William Cullerne Bown stresses, 'The CORE repository, available in bulk, was a breakthrough. Now our algorithms outperform even those from huge publishers'. However, working with commercial partners is not straightforward. 'The definition of commercial vs non-commercial use is creating uncertainty', Knoth explains. 'I don't see it as a problem for the commercial sector - if companies want to do it they will do it - it is just a problem for the university sector, where we are not allowed to take risks.'

CASE STUDY 2: ACTING IN RESEARCHERS' INTERESTS

Tracking experiences under the exception

The Libraries and Archives Copyright Alliance (LACA) is a UK umbrella group convened by CILIP (Chartered Institute of Library and Information Professionals). LACA brings together the UK's major professional organisations and experts representing librarians and archivists to lobby in the UK and Europe about copyright issues which impact delivery of access to knowledge and information by libraries, archives and information services in the digital age.

Following adoption of the UK's copyright exception in 2014, LACA has been working with Universities UK (the representative bodies for executive heads of higher education institutions) to collect evidence of problems encountered by UK researchers trying to use the TDM exception.

Acting on researchers' behalf

In late 2015 LACA was alerted to the case of a UK academic trying to mine a publicly accessible website, which used Captcha technology that prevented him from downloading more than a few records at a time. The researcher contacted the rightsholder to request access, using an online form on their website as no other contact information was available, but did not receive a reply[1].

LACA therefore submitted a complaint to the UK Intellectual Property Office (IPO), seeking a remedy for overly restrictive technological protection measures that prevent permitted acts. The preparation of the complaint took several days, but was seen as a valuable opportunity to clarify the rights of researchers in this area.

Over the period September and October 2015 LACA made further efforts to contact the rightsholder directly, without success. In November 2015, LACA received confirmation from the IPO that the complaint was not within scope of the exception, as 'the [relevant] section [of legislation] does not apply to copyright works made available to the public on agreed contractual terms in such a way that members of the public may access them from a place and at a time individually chosen by them'[1].

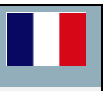
Since users of the website were able to download the databases on-line, the terms and conditions governing access to those works were considered to prevail. As a consequence, the IPO advised that the only option available to the researcher was to approach the owner of the website to request permission to copy/extract the data.

On this occasion, LACA's intervention did not achieve the desired result. However, reflecting on this experience, UK academic librarians from the London School of Economics and the University of Kent concluded: '…It is important to keep fighting battles even if you think it's unlikely you will win. Without further applications to government to address this issue on a case-by-case basis there will be no possibility to change the status quo'[2].

[1] CILIP. Libraries and Archives Copyright Alliance Test Case. (2015)

[2] Secker, J., Morrison, C., Stewart, N. & Horton, L. To boldly go… the librarian's role in text and data mining. CILIP Update Magazine. (2016)

CASE STUDY 3: THE ROLE OF THE NATIONAL LIBRARY



France's national library, the Bibliothèque nationale de France (BnF), has long balanced the needs of researchers against the requirements of copyright law. Emmanuelle Bermès, the BnF's Deputy Director for Services and Networks, sees significant scope for the organisation to support further development of TDM amongst the French research community.

Mining the web

In 2006, the French Heritage Law ('Code du patrimoine') was amended to extend the 'legal deposit' requirement to include the web. Since that time, notes Bermès, the BnF has been archiving French domain websites, in conjunction with INA[3]. Obtaining permission to mine live web content is practically impossible due to the number of rightsholders involved. However, legal deposit allows the BnF to create a rich corpus based on freely-available, archived web content.

The catch is that, like the rest of the BnF's legal deposit collections, the archived material cannot leave the BnF's premises. The parallels with TDM are clear – the data is immensely valuable to researchers, but must be strictly controlled. 'We are already acting as a third party in this case – and creating copies of material for text-mining purposes', explains Bermès.

A learning process

BnF has also gained insight into how best to support researchers in using TDM. Bermès points to the need for technical specialists to work closely with disciplinary experts, and stresses: 'This is not a service that the library can simply deliver, it is more like a partnership'. The BnF's text-mining software has been developed collaboratively with researchers, but there is no 'one size fits all' solution. Some researchers want raw data to mine for themselves, others rely on the library to pre-process data and provide the interface to interrogate it.

Providing assurance

Bermès also points to experience gained by the BnF with the creation of braille and large print versions of published materials for the disabled. 'Publishers didn't want to give source files directly to the associations who'd produce the braille editions,' explains Bermès, 'but without these they cannot create them.' In response, a government decree entrusted the BnF with creating a secure platform for exchanging the files, and establishing a commission to review applications for access. 'This is working really, really well' she observes.

Envisioning a trusted third party for TDM

In considering the BnF's role in facilitating TDM, Bermès zeroes in on a critical question: 'If you project our role into TDM, [it] could be to make sure that requests to conduct non-commercial research are actually research, and are actually non-commercial'. BnF might rely on a commission established by the parties for this purpose, and could also take responsibility for preserving the extracted corpora for future re-use and replication. 'The system would be here to reassure publishers that nothing will happen to the data, and to make sure that people who are undertaking text-mining are allowed to do it', she explains.

[3] Institut national de l'audiovisuel, a repository of French radio and television audiovisual archives.

Looking to the future

Bermès sees value in a process to assess the purpose of research, but cautions that questions of infrastructure and cost must also be considered. A secure, cloud-based service which makes the data available and allows users to run their own instances of search tools is one possibility, but 'we are very far away from that at the moment'.

What is also clear is that no single organisation could field all requests for access to copyrighted material. 'At BnF we wouldn't seek to act as a trusted third party for the big foreign publishers, it's really not our area of activity – this would be ISTEX,' explains Bermès. 'We would handle French language publications, they would have the big foreign publishers - but there would be some overlap.' With the need to accommodate INA's role in audio-visual material, a three-way collaboration may therefore be required. The priority, argues Bermès, is to keep things as simple as possible: 'If we can have a lightweight process, with minimal infrastructure, that makes sure that the spirit of the law is honoured, that would be a good outcome.'

CASE STUDY 4: MINING OF COPYRIGHTED MATERIAL

Our consultation indicates that very few of the researchers employing TDM techniques in France base their research on mining of the scientific literature. This can be attributed to three main factors.

The nature of the research question

First, the selection of material to be mined is a conscious choice dictated by a research question. Amedeo Napoli, Head of the Orpailleur team at LORIA research laboratory in France provides an example: 'We are working on patent mining. We do not necessarily work with big data but also with complex data, and we want to extract in the very best way what's in a document. … We take a more knowledge-based approach.' Patent texts are publicly accessible and can be mined without restrictions.

Legal uncertainty

Second, French researchers often do not know what they are allowed to mine with full legal certainty. Therefore, they turn to material that is freely available on the internet, or make only limited use of scientific literature. Julien Velcin, Associate Professor of Computer Science at the University of Lyon 2, studies representations people build when they come in contact with information. His work relies in part on information from the web sources such as Twitter, blog posts and news sites such as the Huffington Post. Velcin notes: 'The first problem is with Twitter, and determining if the data belong to the company - we are not sure'. He adds: 'if you were to publish a scientific paper, you have to publish your data and your research must be reproducible, but we may not be allowed to do this because of copyright issues'.

Ease of access

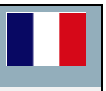
Third, with a lack of enabling legislation and high demands for research output, academics often do not wish to take the risk that data collection could be obstructed by copyright. When it comes to mining scientific articles, Velcin uses sources that freely available on the Internet: 'I work mostly on abstract and title, not on the full publication, because this information is easy to gather from the web site'. In a study of the evolution of the domain of geography, he gravitates towards material available to researchers from the social sciences and humanities web portal Persée[4]: 'We don't want to deal with copyrighted material other than within the Persée project…. we have very strong copyright issues in France, so for us it's very difficult to get this kind of information'.

Change on the horizon

With the TDM exception now in place, new possibilities will emerge for the mining of scientific literature. Napoli observes: 'I think if we have easier access to full papers, for example, it will help us to develop other practices and open doors to different activities that can allow us to have better results'. Unlike in the UK, however, researchers such as Velcin who mine social media sources are unlikely to benefit, since the French exception's scope is restricted to '…textes et de données incluses ou associées aux écrits scientifiques'. For Velcin, this means grey areas still remain: 'Twitter data is somehow public data - but not wholly public. There was no clear ruling on that in French law'.

[4] Please see http://www.persee.fr/ for more information

CASE STUDY 5: REINING IN RESEARCH AMBITIONS



TDM is already firmly established at the French National Institute of Agricultural Research (INRA). 'We have three different teams working in this area', explains Mathieu Andro, INRA's Digitization and Text mining Projects Manager. One of INRA's most successful projects has been the mining of over 30,000 digitised agricultural reports, produced every 15 days since World War 2. With permission to mine the materials granted by the Ministry of Agriculture, INRA has been given free rein to develop a corpus and unlock a wealth of data on pesticides, weather, and disease occurrence. Yet when it comes scientific literature, the picture is much more complex.

What is a domesticated fish?

As worldwide demand for fish continues to rise, farming of aquatic organisms will need to replace the harvesting of wild fish stocks over time. To help develop strategies for managing this, INRA initiated a 2012 project to classify fish species according to their potential for domestication. It's the sort of question that is tailor-made for text-mining, and Mathieu and his team were brought in to analyse the literature.

Having identified 60,000 relevant papers, they embarked on a lengthy process of seeking publisher permission to mine them. 'They usually answered yes', explains Andro, 'but it takes a lot of time and it is difficult to download so many texts because [even when we had permission] the publishers' servers frequently blocked us'. Andro estimates that it was a week's work just to secure the licences, but publishers' technical protection measures presented a much greater barrier. Even with a team of people working to manually download the articles, and the aid of automated tools such as Endnote, the original goal of mining 60,000 articles proved an impossible task.

Instead, he turned his attention to the literature available in ISTEX. 'It was the best solution for us - but you don't have all the journals, particularly the most recent ones, and the quality of content is not always very good', he cautions. Although all content in ISTEX is normalised (using MODS and TEI standards), the texts themselves sometimes suffer from poor optical character recognition (OCR), and issues with structure and tables. 'We had to abandon the list of 60,000 articles, because we knew we wouldn't find everything in ISTEX', says Andro, 'but it was the only viable option in the circumstances'.

Prospects for the future

Mathieu Andro is ambivalent about the potential benefits of legislative change. 'The ability to download texts without seeking authorisation would be very welcome', he acknowledges, 'but the real problem is getting the data from the publisher servers'. He has high hopes for ISTEX as a text-mining resource, but also points out that gold open access may offer the best long-term solution to the access issue. Whatever approach is adopted, he stresses that preserving the data used in TDM is paramount: 'If we destroy the corpus, we would soon find there is a different question to ask on the same subject - but we would have to start all over again'.

CASE STUDY 6: 'PUBLISHERS STOPPED ME DOING MY RESEARCH'

The Meta-Research Center at Tilburg University in the Netherlands uses statistical tools to detect data fabrication (among other things) and support the reliability of scientific literature. Chris Hartgerink, a PhD student at the Center, uses tools developed by Contentmine in the UK to search thousands of Psychology articles for patterns in data and text that may be associated with fabrication. However, gaining access to the articles he needs has proven far from simple.

'Elsevier stopped me doing my research'

In late 2015, Hartgerink posted a blog outlining the difficulties he had encountered in downloading papers from Elsevier's ScienceDirect website. The blog was widely shared on social media and reported in Nature[5] - but the story did not end there. In subsequent months, Hartgerink also found his access blocked by Wiley[6], and estimates that almost a third of his time in early 2016 was spent 'learning about the copyright problems, reading terms and conditions, meeting with management, discussing problems… that's time that would otherwise be spent doing the research'. The end result was that Hartgerink was unable to achieve his original goals: 'I scaled down my TDM research, and had to exclude two publishers, meaning I was able to do some of the research, but not what I set out to do'.

Assessing the UK exception

Given his experiences in the Netherlands, Hartgerink sees tremendous value in a copyright exception: 'What the UK exception does is empower researchers to do the TDM research they want to do'. He contrasts this with the position in much of continental Europe, where the onus is on researchers to make the case to institutional management and publishers.

Hartgerink is now exploring scope to undertake his research with the Contentmine team in Cambridge: 'I know I can go to the UK and download these papers … because I am then geographically relocated I am allowed to do it.' Nevertheless, he is not surprised that most UK researchers have been slow to take advantage: 'People have to become aware, have to become convinced that the legal exception is covering all of what they are doing… it's a very complex subject and it's very easy to get [the] wrong idea. This is part of the reason why uptake is low'. Contentmine has now offered him a 'safe haven', but Hartgerink stresses that this is not a systemic solution.

Maximising the impact of TDM

Hartgerink sees huge potential for TDM to speed up research, but only if access to content can be improved. He points to the freely accessible APIs offered by PLoS, eLife and PeerJ as a model for publishers to follow, and also sees potential in CrossREF's TDM service: 'For those publishers that don't apply additional conditions it works pretty great'.

Hartgerink also voices concerns about the granularity of the publishing market, given scientific articles are published by hundreds of different publishers. In order for TDM research to cover all of these, it is essential that publishers adhere to open standards that simplify the processing of content from different platforms in a reliable manner. Currently, different APIs by different publishers

[5] Hartgerink, C. H. J. Elsevier stopped me doing my research. (2015)
[6] Hartgerink, C. H. J. Wiley also stopped me doing my research. (2016)

contain documents that are built up in different ways and require different software to properly process them.

He remains worried, too, that few researchers have the necessary legal expertise to make fine-grained judgements between commercial and non-commercial research. Achieving clarity on this point is critical if they are to be encouraged and empowered to do TDM, he argues. For Hartgerink, the best way of achieving this is for an exception to cover research as a whole: 'The added value of TDM research is immense, but if it is only for non-commercial research then the share [of that value] is small. Research is of great value to society, regardless of whether it is done in universities, in commercial companies, or by citizens who do it as a hobby'.

CASE STUDY 7: SOLVING THE MANY-TO-MANY PROBLEM

The potential for a trusted third party to facilitate TDM on behalf of publishers and end-users has long been recognised. A number of solutions have emerged from within the publishing industry itself, often in response to demand from the corporate market. One example is the RightFind for XML service, developed by Copyright Clearance Center, Inc. (CCC), a U.S. company that provides collective copyright licencing services for corporate and academic users of copyrighted materials.

Defining the problem

'Having a right to mine and being able to mine effectively are two different things' explains Roy Kaufman, CCC's Managing Director, New Ventures. 'People are asking whether they have a right to mine, but the question should be "how I can do it?"'.

CCC's primary market is the mining of medical literature under licence, and in this area the quality of content is key. 'There is a difference between scraping (low fidelity, giant blob of text, no metadata) and text-mining' says Kaufman. 'For our users who spending a lot of money on content, they want high fidelity XML.'

In this respect, corporate text-miners are in a privileged position compared with public researchers - but both face the challenge of sourcing content from multiple sources. This is where a trusted third party can help, he argues: 'It's about not having to go to every publisher, not having to scrape, not having to get a feed. Everything takes time, so aggregation is really important'.

The right tools for the right job

Text-mining in academia shares some challenges with the corporate world, then, but there are other notable differences. One relates to the software tools used, according to Kaufman: 'All pharmaceutical companies are text-mining, sometimes the tools are customised, [but] they use the same technologies'. By contrast, the software tools used in academia are more diverse: 'If I am talking to academic institutions, librarians, it's not clear what tools they are using. Biomedical miners use very different tools from cancer miners'.

The danger is that this proliferation of tools and techniques inhibits researchers' ability to get the most out of their content. 'If I were government or library trying to enable mining', observes Kaufman, 'the question would be: how good are my tools?'

Lessons learned

CCC's experience, and the emergence of other providers of similar services, shows that there is significant scope to streamline the text-mining process. 'I would be the last to say you couldn't do a trusted third party option, because that is what we are,' stresses Kaufman. However, he warns of the dangers of underestimating the challenge involved. 'You need XML that is standardised across and within the publishers, and that costs money,' he notes. He also points to concerns over data protection and security: 'It's hugely important to publishers that [their data] is not released into the wild'.

Both publishers and content users are likely to ask searching questions about the credibility of any intermediary, the contractual terms applied, and their ability to deliver a viable service. All this is surmountable, but it is not something to be embarked upon lightly, concludes Kaufman: 'There's a lot of work - we've done it, so we know'.

CASE STUDY 8: A CONTINUED ROLE FOR LICENCING

The adoption of a copyright exception is intended to reduce transaction costs, by eliminating the need to negotiate access on a case-by-case basis. In practice, publishers' rights to use technical protection measures under UK law means that disputes over access terms can persist even under an exception.

Working with publishers

Patricia Killiard is Acting Deputy Director, Academic Services at Cambridge University Library, and oversees licensing of scholarly content on behalf of the University. 'We've had some [licensing] examples that have been really urgent around full-text humanities content' she recalls. She cites the case of a US-based publisher, where researchers needed to secure access to licensed content as part of a research grant: 'The publisher sent them a form to sign, which basically meant trying to tie down what they could do, in complete opposition to the TDM exception'. The proposed terms meant that only single copies of the content could be made, and each individual project and researcher would need their own licence to obtain a copy. While the university stressed that researchers were entitled to mine the content as a result of the exception, the publisher argued that they were entitled to take 'reasonable steps to protect their content' - even though the UK TDM exception explicitly excludes the possibility that it can be overridden by contract.

Finding a way forward

While this was an extreme example, it illustrates the continued difficulties faced by researchers seeking to mine content under the copyright exception. Killiard believes consortial bodies such as Jisc Collections in the UK and the Couperin Consortium in France have an important role to play in breaking the impasse over access. These bodies work with publisher trade associations to agree model licence terms, and the adoption of clauses covering TDM could support the practical implementation of a copyright exception. 'Licences allow you to do things that you can't do under the law', Killiard explains. 'The law permits publishers to apply technical protection measures, so there has to be some negotiation about what publishers need to do to protect their content'. Ultimately, though, the library community may need to be prepared to test matters in court. So far, says Killard, 'We haven't been brave enough, we haven't really tested these exceptions legally'.

CASE STUDY 9: GATE – SOFTWARE AND TOOLS FOR TEXT-MINING

GATE is a large framework of open source software for text mining whose aim is to develop tools and methods to make text mining as easy as possible. GATE works by collaborating with universities, research labs, SMEs and corporations who have large collections of texts, but lack text mining capabilities to extract useful information from them. GATE contributes its text mining tools and expertise to projects in areas such as cancer research, drug research, decision support, recruitment, annotations and social media.

Sharing text mining tools, training and expertise

Mark Greenwood, a member of GATE group, explains the role of GATE in a basic text mining process: '…Every time you have a new project or data source that you want to process you hit issues about how the documents are structured, oddities of formatting, etc. When we load documents into GATE, … most of the tools convert the source material into something that looks a bit like HTML. We rely on a lot of tools, including one that is designed to allow API access to documents in various formats – Word, PDF, Excel'. GATE tools allow extracting text components and formatting the text, which creates a good foundation for further mining steps and answering specific research questions.

An advantage of GATE's frameworks is that it processes documents from different sources and inter-operates with other people's tools. It gives a user an ability to pick and choose tools and run them in sequence to achieve what they want. GATE brings in software to create a corpus, map documents on a concept tree and frame search on the concept tree. 'But your users don't care. They're just happy because now they can find stuff'[7].

What changed after the UK copyright exception for TDM

GATE group has existed for over 15 years doing TDM long before the exception. Its tools have been downloaded about 300 000 times since 2005[8]. One area where the exception made a difference for GATE is social media. Greenwood explains: 'A lot of the platforms specifically prohibit you from doing any large-scale processing of content produced by users… With the exception, they can't stop us, because they have been published publicly'. Overall, the exception has given GATE confidence to mine publicly-available user data and made initial research much easier.

Concerns

The interpretation of the non-commercial provision remains a concern. GATE researchers remain unsure whether they are allowed to make derivative work, such as a machine learning model, available for commercial applications, or if international collaborators may use research done under the UK exception for commercial purposes. 'The problem with the exception is it's not been tested, and that's pretty much always the problem with laws relating to technology', concludes Greenwood.

[7] Please see https://gate.ac.uk/2mins.html for more information
[8] Please see https://sourceforge.net/projects/gate/files/gate/stats/timeline?dates=2015-08-26+to+2016-08-26 for more information

CASE STUDY 10: ISTEX – AN INVESTMENT IN THE FUTURE

The ISTEX project is a vast programme for the acquisition of scientific resources, aimed at setting up a digital library for the benefit of members of higher education and research establishments across France. Since 2012, the French National Research Agency (ANR), the State and the CNRS have invested 60 million Euros to acquire resources and set up the platform. Inspired by a German initiative[9], ISTEX is nevertheless the first attempt to group together several million multidisciplinary and multilingual documents in a normalised format.

A licensing-based approach

The potential such a vast resource holds for text miners was recognised from the project's outset. Accordingly, the right for beneficiaries and authorised users to mine data was included in the national licences negotiated with the various content providers. Under the terms of these licences, this right is granted provided the text mining activity is 'in accordance with the beneficiaries' mission'.

Jean-Marie Pierrel, Director of ATILF[10] at the University of Lorraine, is responsible for the provision of research and services for ISTEX. He explains, 'An exception would not change much for ISTEX users, as the contracts are already signed. However, it would have avoided protracted negotiations over the existing licences, and would greatly facilitate the integration of new resources'.

A fully-functioning infrastructure

ISTEX comprises both hardware and software infrastructure designed to get the most out of the acquired resources. Publisher data is grouped into a single, normalised corpus, allowing researchers to mine thematic sub-corpuses, regardless of their origin. Researchers can also request the right to download sub-corpuses from ISTEX and use their own tools to mine the material.

Any member of a French higher education or research establishment is able to benefit from ISTEX's wide range of software tools and powerful search engine. The project budget also supports a dedicated team responsible for pre-processing, standardisation and enrichment of data and metadata. Meanwhile, the preservation of technical copies within ISTEX is assured by the Institute for Scientific and Technical Information (INIST)[11]. In many respects, the service is a text miner's dream come true.

No panacea

It is clear that ISTEX is an invaluable resource for French researchers, but it is not a panacea. ISTEX only holds older archives at present, with most of its content published prior to 2010. Negotiations are underway to acquire more recent content, but the ISTEX corpus remains only a fraction of the scholarly literature as a whole. Where their research questions can be answered within the confines of this corpus, ISTEX offers French researchers a huge advantage over their counterparts in the UK, or elsewhere. Where the material to be mined is of more recent origin, or comes from other sources, researchers in both countries face common challenges.

[9] Please see http://www.nationallizenzen.de/ for more information
[10] Analyse et Traitement Informatique de la Langue Française
[11] Please see http://www.inist.fr/ for more information

CASE STUDY 11: CONTENTMINE

ContentMine is a not-for-profit company in Cambridge, UK, whose declared mission is to liberate scientific facts from academic journals and enable anyone to perform research using content mining techniques. It started in 2014 when Peter Murray-Rust, the Cambridge University Professor Emeritus of Chemistry and an advocate of open science, sought to use the UK exception to mine the scientific literature as a new form of research. His idea resonated with the aims of Shuttleworth Foundation, a privately-funded 'Purpose Trust' that supports people 'who are unafraid to re-imagine the world'. The Foundation committed $350,000 (€310,000) to enable Murray-Rust to lead the ContentMine project to demonstrate the usefulness of content mining and to provide tools and services to others.

'ContentMine finds the facts, so that you don't have to'

ContentMine are academics, software developers and managers who develop mining software and show how to use it to extract facts from research articles. The web portal contains detailed instructions on how to create a corpus and run the right mining plug-ins to get papers, normalise them and process them to search for key terms and more. It is often uncharted waters. 'There are no TDM standards – ContentMine is pioneering all of this', says Murray-Rust. All ContentMine software is open source available on GitHub.

Holistic support

ContentMine are also ardent advocates of TDM, and are committed to demonstrating what is possible, both technically and legally. 'I hope to show that knowledge should be for everybody, not just for very rich universities who use tax payers' money to pay publishers to prevent the spread of knowledge', explains Murray-Rust.

The company also provides training and support in content mining. It has delivered around 300 workshops and its work is widely promoted through videos[12] and presentations[13]. Talks by the ContentMine team have reached an estimated audience of 2,000 people.

Progress and impact

Two years that the company existed has not been long enough to demonstrate dramatic impact. 'It's not easy to come up with massive conclusions, because we are developing tools at the moment', recognises Murray-Rust. The most ground-breaking outcome so far has been creating a phylogenetic tree from taxonomy literature[14]. As the tools are maturing, the company is about to embark upon more subject-specific research. Six young talented international researchers, including one based in France, have recently become the first ContentMine fellows. The fellows will apply mining techniques in fields such as computational and systems biology, neuroscience, oncology and taxonomy. The ContentMine portal has also about 30 active users. The ContentMine software users and fellows are early adopters who want to use TDM to fast-forward their research, and accelerate its impact on society.

[12] Discovery of research papers relevant to the zika virus outbreak
[13] Cochrane workshop 2016
[14] Please see https://blogs.ch.cam.ac.uk/pmr/2014/06/25/content-mining-we-can-now-mine-images-of-phylogenetic-trees-and-more/ for more information

CASE STUDY 12: TRAINING THE NEXT GENERATION OF RESEARCHERS

The GREYC Laboratory is a research unit placed under the joint responsibility of CNRS/University of Caen Basse-Normandie and ENSICAEN. With over 220 members, its work covers research and academic skills in the fields of computer science, electronic and electrical engineering in Basse-Normandie. The laboratory places great importance on technology transfer and research training of junior researchers.

Dealing with data overload

François Rioult, a faculty member at GREYC, is responsible for equipping students and researchers with the skills needed to make effective use of TDM. 'I'm particularly involved in the fundamentals of data mining, the mathematical aspects', he explains, 'but my work also has lots of applications related to genomics, sport, text mining, video mining, health, and so on'.

Rioult notes that a key challenge of data mining is the sheer volume of knowledge that can be created: 'The first issue is that we need experts to make sense of all this knowledge – and this is a real difficulty, because our algorithms can produce too many results.' He argues that data scientists therefore have a responsibility to present knowledge to domain experts in the right way, to make best use of their time and expertise.

Rioult also emphasises that 'the data scientist has to be present from the very beginning of the process to the very end, and to master every part of every stage.' It's a hard job, he concludes, and as a result many data scientists struggle to present their results effectively to a domain specialist.

Asking the right questions

Conversely, researchers without expertise in TDM may not know how to take advantage of it: 'I have lots of people who come to see me because I have some expertise on data mining, and they have lots of data, but they don't know what to do with them. They have no questions'. It is clear to Rioult that TDM must be combined with specialist knowledge of the subject matter to be effective: 'We have algorithms to answer questions, but we do not have algorithms to ask questions,' he notes.

Skills gaps are certainly part of the picture, with Rioult stressing that 'there are lots of job opportunities' for data scientists. However, the research community as a whole needs to reconsider its approach in light of the new opportunities presented by big data and TDM. 'The world is now waking up to the fact we have lots of data', notes Rioult, 'but people continue to ask [research] questions from the previous century'.

CASE STUDY 13: ENABLING TDM AT THE UNIVERSITY OF CAMBRIDGE

The Office of Scholarly Communication (OSC) at the University of Cambridge was formed in 2015 to help the University keep abreast of new scholarly communication tools, techniques, policies and practices. A joint initiative of the Library and Research Office, it supports researchers and administrative staff across the institution, and has a key role in envisioning the future role of Cambridge University Library.

Limited demand

Uptake of the copyright exception by Cambridge researchers remains limited, according to Dr Danny Kingsley, Head of the OSC. While initiatives like ContentMine are leading the way, the majority of researchers are either unaware of the exception's existence, or unclear how it could benefit their research. 'This doesn't mean it's not happening at all', cautions Kingsley, 'but the research community is not thinking "we need to go to the library to talk about this"'. She also points to the fragmented international copyright environment as a limiting factor. Researchers need to share text-mining data and corpora freely across national boundaries, and several have expressed concerns over the implications of this for their European partners. The hope was that the UK's exception would give its researchers an advantage in this area. In practice, many are inhibited by the lack of a similar exception in the rest of Europe.

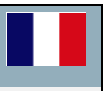
Rethinking the library's role

Where researchers do seek assistance with TDM, the issues involved can prove complex and time-consuming to resolve. 'We've had to work through who is responsible for this kind of thing in the library', Kingsley explains, noting that 'there's often a misunderstanding of what is involved in TDM, and what kinds of skills and knowledge are needed'. The development of these skills is part of a broader need for the library community to reshape itself for the digital age. 'Librarians will say that they have knowledge of copyright, but what they mean is not what we are after', Kingsley argues. 'The kinds of skills and knowledge that are needed now - it's a whole new world'.

Developing a TDM support service

To date, the focus at Cambridge has been on handling TDM licensing queries more effectively, but Kingsley sees scope for librarians to provide more technical support in future. Libraries will also need to be proactive in managing risk to the institution, representing researchers' in negotiating access to content, and developing an institutional position on the use of TDM. Much of this comes down to providing guidance to researchers on what TDM is, and identifying named people in the library who can support them. The aim, according to Kingsley, is to get to the point where the library can say, 'If you've got a question about TDM, come to us'.

CASE STUDY 14: MEETING THE COST OF CONTENT



The Complex Systems Institute of Paris Ile-de-France (ISC-PIF) is dedicated to the development of innovative and interdisciplinary research on complex systems. David Chavalarias, the Institute's Director set out 10 years ago with an ambitious goal: 'to build a picture of science and its evolution over time'. The Institute uses large scale text-mining as a key means of fulfilling this goal.

The cost of access

In 2015, the Institute sought to map the evolution of climate change science for an exhibition at Cité des Sciences et de l'Industrie, Paris. 'This required new access rights to be able to mine our WoS database the proper way and display titles of relevant papers online. Again, we had to negotiate with Thomson Reuter, for this specific right'. Chavalarias points not only to the time taken to regularly negotiate access and the cost, but also the fact that, since access depends on the good will of the providers, it be cancelled any time. This creates the risk of ruining long term scientific projects: 'You just can't do sustainable TDM if you don't buy the database at full cost.'

Opening up access

The cost of gaining access to content remains a significant barrier to greater use of TDM, according to Chavalarias. He hopes, though, that a copyright exception will open up more possibilities once this access is secured: 'Maybe we'll be able to do more data mining… with less restrictions, since we are not commercial'.

Like other French researchers, he also sees real promise in ISTEX: 'What they are doing at ISTEX is really wonderful, because they are doing things properly on the data they handle, and this is really helpful'. At present, ISTEX offers only a fraction of the articles indexed in Web of Science or Elsevier's Scopus database, and lacks content from recent years. The hope, according to Chavalarias, is that the new law will provide the impetus for its coverage to be further expanded.

Realising downstream benefits

The open-source software developed by ICS-PIF to track the evolution of science is well-suited to identifying other emergent topics, trends and debates. 'The method is completely agnostic in terms of what kind of documents you have,' explains Chavalarias. It has already found applications within the not-for-profit Pasteur Foundation and the European Commission. Chavalarias also sees potential for a wide range of commercial applications: 'A company can take data on, for example, all their patents, and apply our software and text-mining techniques to that', he notes.

Yet making data and tools available for others to re-use is not always straightforward. Chavalarias notes a number of areas of uncertainty faced by researchers: 'Can I process this data? Can I put it online for non-commercial use? Should I control who has access, or should it be open to the public?' Addressing these concerns and facilitating collaboration between researchers are critical to maximising the benefits of TDM.

CASE STUDY 15: CULTURAL BARRIERS TO TDM

A postdoctoral researcher at the University of Cambridge, Ross Mounce has used text-mining across a range of projects in the fields of palaeontology and phylogenetics, both in academia and at London's Natural History Museum (NHM).

A moment of liberation

Reflecting on the period prior to the 2014 exception, Mounce observes 'my work was self-limited because I didn't want to get in trouble. I had previous experience of being blocked by publishers, and because of those experiences I didn't want to push the boundaries too much'. That all changed from 1 April 2014: 'As soon as the UK legislation came in that emboldened me, because I knew I could legitimately download a much greater quantity of material'. Over time, the exception has allowed him to build up a rich corpus of hundreds of thousands of papers in biodiversity sciences, and mine them for specimen identifiers.

Pinpointing the benefits

For Mounce, the benefits of the exception cannot simply be measured in person-hours saved. 'It's really binary,' he explains, 'I could try to give you an estimate for how long it would take to [individually] download 10,000 papers, but that would be silly - I just wouldn't have done it'. He points instead to the benefits derived from his work, particularly in helping the Natural History Museum measure its impact on research. 'You can have papers that use a thousand different specimens, all uniquely identified', he notes, 'but the museum never gets to know about that, unless someone does some text mining'. By linking specimen identifiers back to the NHM's own catalogue, Mounce was able to demonstrate exactly how each type of specimen in its collection was being used. This informs NHM's internal decision-making processes, and helps it identify the most valuable parts of its collection. Mounce himself has benefited in career terms, attributing his postdoctoral position at Cambridge to the skills acquired through his work on text-mining.

A reality check

While he remains a strong advocate of the UK's exception, Mounce is also clear on its limitations. 'The restriction to non-commercial use has been a huge hindrance to my work', he observes. 'If I was allowed to sell this as a service to institutions, to go round museums and tell them which papers your species has been mentioned in, I could develop a start-up around that. But that isn't going to happen in the UK'. He also points to a lack of training, skills and incentives, as well as cultural resistance to text-mining in many parts of academia. 'There are so many obstructions in the way of doing this research and doing it well, it is just too hard and so people do other things,' he concludes. 'I am in that position myself - there are no budgets out there, it is not even on the radar that this is possible'.