A New Declaration of Rights: Open Content Mining

In a recent investment report, analyst Claudio Aspesi concluded that a new front had opened up in the Open Access (OA) debate. Writing in April, Aspesi noted that academics are "increasingly protesting the limitations to the usage of the information and data contained in the articles published through subscription models, and — in particular — to the practice of text mining articles." Aspesi is right, and a central figure in this battleground is University of Cambridge chemist Peter Murray-Rust. A long-time advocate for open data, Murray-Rust is now spearheading an initiative to draft a "Content Mining Declaration". What is the background to this?

(There is a short Q&A with Murray-Rust at the end of this text. Click here to go directly to it)

When I interviewed Peter Murray-Rust in 2008, he expressed considerable frustration at the difficulties he was experiencing in trying to extract and reuse the data published in scholarly journals — even where his university had paid an electronic licence to access the content.

What Murray-Rust wanted to do, he explained, was to capture the "embedded data" contained in the tables, charts, and images published in science papers, along with the "supplemental information" that often accompanies papers. To do this, he had developed a variety of software tools to mine large quantities of digital text. Having extracted the data he then wanted to aggregate them, compare them, input them into programs, use them to create predictive models, and reuse them in a variety of other ways.

However, he was having huge problems achieving this, not because of any technical issue, but because of uncertainty over copyright and publishers' insistence that a licence to read journals does not encompass the right to mine them with software.

To add to Murray-Rust's frustration, many of his colleagues were either unsympathetic or uncomprehending. Even more galling, the Open Access movement — which should have been a natural ally — was more interested in making papers freely available to eyeballs, than to software. Even papers published in OA journals, he noted, are often released under licences that do not come with reuse rights.

In pursuit of his dream, Murray-Rust became a formative voice in the creation of the open data movement. Open data, Murray-Rust explained to me in 2008, is data "free of any restraint on access and on reuse." Recently, however, governments have tended to lead the way in urging for open data, spawning a generation of data wranglers; open scientific information has often lagged behind, but is now beginning to be seen as a central issue.

Four years later, Murray-Rust is still frustrated. He is not, however, a man to give up, and he continues his advocacy today under the rubric of "open content mining". Essentially, this is text mining ***plus***. As Murray-Rust explains today, he views the mining of scholarly journals as a hierarchical activity, with content mining encompassing not just the mining of text and data, but other types of content too, including images, tables, graphs, audio, and video.

Simply using the term "text mining", he adds, "might imply that anything other than text should be protected by the 'content provider'. However, I and others can extract factual information from a wide range of material."

The good news is that the research community is finally beginning to understand what Murray-Rust has been "banging on about" for all these years, as are research funders and governments, and Murray-Rust believes the door to what he wants is finally beginning to open.

However, he says, it is imperative that text mining advocates push hard at that open door if they want to achieve their objectives. To this end, Murray-Rust recently convened an ad hoc group of interested parties to draft what he calls a "Content Mining Declaration" (disclosure: I am a member of the group).

Copyright and contracts

So what has changed since I spoke to Murray-Rust four years ago? So far as publishers are concerned, he says, very little has changed. There have, however, been a number of broader developments that are seeing Murray-Rust's call for content mining rights begin to get a more receptive hearing.

First, there is a growing acceptance that traditional IPR is impeding or preventing a good deal of innovation in today's digital environment. This has made governments more open to the suggestion that it may be necessary to recalibrate copyright for the networked world.

In November 2010, for instance, the UK Prime Minister David Cameron commissioned Professor Ian Hargreaves to review the current situation. This led to the publication in May of last year of a report — *Digital Opportunity: A Review of Intellectual Property and Growth* — in which 10 major changes to the current intellectual property regime were proposed, including changes to copyright laws that Hargreaves concluded "obstruct innovation and economic growth in the UK". If these are all implemented, *The Guardian* suggested last year, it will amount to an "overhaul of copyright laws" in the UK.

Importantly, one of Hargreaves' recommendations was for the UK government to "introduce a UK exception … under the non-commercial research heading to allow use of analytics for non-commercial use … as well as promoting at EU level an exception to support text mining and data analytics for commercial use."

In arguing the case for this exception, Hargreaves cited an example that highlights both the potential benefits and the current obstacles to text mining scholarly papers, and demonstrates why Murray-Rust is so frustrated.

Specifically, Hargreaves explained how the Mahidol-Oxford Tropical Medicine Research Unit, based in Thailand and supported by the Wellcome Trust, wanted to make a set of around 1,000 journal papers from the first half of the twentieth century freely available to researchers.

It would, explained Hargreaves, have been "a unique and unrepeatable experiment" as the papers both describe malaria in indigenous peoples and soldiers, and give details of the malaria therapy that was being used at the time. As such, the papers offered "potentially significant insights for the development of methods for preventing and treating malaria today."

Unfortunately, Hargreaves noted, it is "often impossible to establish who are the copyright holders in these articles, many of which appeared in long defunct journals — they are orphan works. Copying them to make them generally available in online form would break the law. Reproducing individual illustrations and diagrams in articles is not possible."

However, stressed Hargreaves, this was not just an orphan works problem. Even if the rights holders were known, he explained, it would "still not be possible to text mine them — copy the articles in order to run software seeking patterns and associations which would assist researchers — without permission from the copyright holders who ***can*** be found, since there is no exception covering text mining."

And even if all these obstacles were overcome, he added, it "would not guarantee that text mining would be possible in future cases" since publishers often impose limitations on text mining in their licensing contracts.

In short, concluded Hargreaves, "Text mining is one current example of a new technology which copyright should not inhibit, but does. It appears that the current non-commercial research 'Fair Dealing' exception in UK law will not cover use of these tools under the current interpretation of 'Fair Dealing'. In any event text mining of databases is often excluded by the contract for accessing the database."

For this reason, said Hargreaves, "any new text mining exception [would also need to] include provision to override any attempt to set it aside in the words of a contract."

Hargreaves recommendations have yet to be implemented, although in August 2011 the UK government accepted all ten of them. When a text-mining exception might appear, however, is unclear. In any case, points out Murray-Rust, it would only apply in the UK. Science, by contrast, is a global endeavour.

Nevertheless, Hargreaves' articulation of the problem faced by text miners was extremely helpful in itself, since it has served to focus other minds on the issue, and encouraged further research into the problem.

On 14th March, for instance, the UK's Joint Information Systems Committee (JISC) published a report on text mining. This listed a number of benefits that text mining could be expected to provide, including, "increased researcher efficiency; unlocking hidden information and developing new knowledge; exploring new horizons; improved research and evidence base; and improving the research process and quality. Broader economic and societal benefits include cost savings and productivity gains, innovative new service development, new business models and new medical treatments."

Unsurprisingly, JISC concluded by reiterating Hargreaves call for a text mining exception.

Massive growth in digital information

The second development that has helped to focus minds on text mining is the explosion in the quantity of digital data. As OA advocate John Wilbanks put it recently on the Open Knowledge Foundation blog, "Data is entering the world at a rate that is so fast it's almost incomprehensible to human brains. It's like trying to comprehend geologic time. The cost of generating data is so low in so many spaces, and dropping like a stone in so many others, that the real challenge is to do interesting things with it."

True, much of the data Wilbanks refers to is not scientific data. And where it is, it tends to be the massive datasets being created in fields like genomics, physics and astronomy, not data that has been mined from scholarly papers. But growing awareness of the potential value that can be extracted from data is all grist to Murray-Rust's mill.

Consider also that 1.5 million new articles are published each year. Processing all this information effectively is going to require more than eyeballs alone. Moreover, while today an archive like the US National Library of Medicine's free digital database of scientific literature in the biomedical and life sciences — PubMed Central — holds only 2.4 million articles (around 11% of the total published biomedical literature), it is growing rapidly; and this growth is not only due to the deposit of newly published articles, but large back files of journal content are being loaded too (including at some point, malaria researchers doubtless hope, the 1,000 papers that Hargreaves drew attention to). Either way, the corpus of research papers will continue to grow rapidly, and people will increasingly want to mine the papers.

It is important to note that the economic benefits of text and data mining have become key in advocating for open data and for text mining.

Last year, for instance, the management consultancy McKinsey & Company published a report in which it estimated that billions of dollars in value could be realised from data. The report was focused on *data* mining rather than *text* mining, and on Big Data at that (i.e. "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyse"). Nevertheless, the growing emphasis on the economic value of data clearly aids the case for text mining. In accepting the Hargreaves recommendations, for instance, the UK government noted that doing so would provide a potential benefit to the UK economy of up to £7.9 billion.

The third development to note is the trend in recent years for launching open government initiatives. Since open government assumes that citizens have the right to access the documents

and proceedings of the government to allow for effective public oversight it means that more and more government data is being made freely available, and not just for transparency purposes, but to enable companies and citizens to mine it and reuse it to create new information, new products, and new value.

And the calls to release government data have a natural affinity with the emerging principle that all publicly-funded information should be available for public scrutiny, including publicly-funded research papers and associated data. This promises a twofold benefit: ensuring that citizens can access information that they have funded, and enabling any hidden value in that data to be discovered and exploited. Thus, as a result of the America COMPETES Reauthorisation Act of 2010, at the end of last year the US Office of Science & Technology Policy (OSTP) issued two Requests for Information on Public Access to Digital Data and Scientific Publications.

The first question in the RFI on Public Access to Digital Data was, "What specific Federal policies would encourage public access to and the preservation of broadly valuable digital data resulting from federally funded scientific research, to grow the US economy and improve the productivity of the American scientific enterprise?"

While neither OSTP RFI specifically mentions text mining, many of the respondents will doubtless have done so. In its submission to the RFI on OA publications, for instance, Harvard pointed out that "[p]ublic access not only facilitates innovation in research-driven industries such as medicine and manufacturing. It stimulates the growth of a new industry adding value to the newly accessible research itself. This new industry includes search, current awareness, impact measurement, data integration, citation linking, ***text and data mining***, translation, indexing, organizing, recommending, and summarizing." (My emphasis).

In its submission to the RFI on data, Elsevier also mentioned text mining, although in reading it one might conclude that it did so more in order to promote the value it believes it can bring to the process, not the value of the data itself. As Elsevier put it, "Federal agencies should also adopt policies that encourage publishers to continue to invest in their journals and in the development of discovery tools for data. For example our article linking tools facilitate entity text-mining (e.g. Arabidopsis Viewer) …"

As such, the suspicion must be that Elsevier is keen to ensure that publishers continue to "own" the data. This not only holds out the possibility of being able to sell back to the research community its own data (in addition to subscription access to the text), but also of licensing intermediary tools to researchers to assist them mine that information. As we shall see, however, researchers would prefer to use their own tools.

Publisher response

Unsurprisingly, therefore, publishers are not happy at the way in which discussions about text mining have been unfolding. Responding to Hargreaves' call for a text-mining exception last November, for instance, the CEO of the Publishers Association Richard Mollett insisted that denying publishers the ability to control access to research papers (and thus the data within them) by allowing permission-free text mining would lead to chaos and, rather than providing economic benefits, would have deleterious financial consequences.

First, argued Mollett, publishers' platforms "would collapse under the technological weight of crawler-bots."

Second, he said, it would impose a significant commercial risk on publishers. "It is all very well allowing a researcher to access and copy content to mine if they are, indeed, a researcher. But what if they are not? What if their intention is to copy the work for a directly competing-use; what if they have the intention of copying the work and then infringing the copyright in it?"

Third, Mollett said, if Britain were the only country to provide such an exception it would put itself at a competitive disadvantage to the rest of the world. "Why run the risk of publishing in the UK,

which opens its data up to any Tom, Dick & Harry, not to mention the attendant technical and commercial risks, if there are other countries which take a more responsible attitude."

In any case, publishers argue, very few researchers want to text mine scholarly papers. Removing control from publishers would therefore introduce serious technical and financial risks for little obvious gain. Commenting in May in *The Guardian*, for instance, Graham Taylor of the Publishers Association said, "Mining requests so far are relatively few and permission is generally willingly and easily granted for non-commercial purposes (see here for the evidence for this statement)."

We will come back to the question of publishers' willingness in a minute, but we should note that the claim that few researchers currently mine journals does appear to be accurate. On 2nd March, for instance, Casey Bergman, a researcher in the Faculty of Life Sciences at the University of Manchester, posted a note on his blog wondering why so few efforts are being made to text mine PubMed Central today.

"Surprisingly, I found that after a decade of existence only ~15 articles have ever been published that have used the entire open-access subset of PMC for text-mining research," he wrote. "In other words, less than 2 research articles per year are being published that actually use the open-access contents of PubMed Central for large-scale data mining or service provision. I find the lack of uptake of PMC by text-mining researchers to be rather astonishing, considering it is an incredibly rich archive of the combined output of thousands of scientists worldwide."

Writing on 7th March, *Nature* summed up the current situation in this way: "Publishers point out that they receive few text-mining requests, so the field can't be very hot. So unless text-miners start to make full use of the content that is available, and request more access to published content — while always being clear about how their project will benefit science — the unsatisfactory impasse will continue."

The JISC report, however, concluded that text mining is currently rare not because there is a lack of interest in doing so, but because publishers take an overly proprietorial attitude to the papers they publish. "[T]ext mining is currently extremely limited within UKFHE," the report noted, "in part at least due to the current licensing arrangements. A text mining exception, if it were to be implemented, would remove a key barrier thus better enabling service solutions supporting text mining to emerge from the market."

We noted Taylor's claim that publishers are happy to allow researchers to text mine. Wiley-Blackwell's Bob Campbell has made the same claim. In an email to Murray-Rust in March, Campbell said, "[A]nyone interested in mining our journal content should contact us. Any such inquiries will be treated on a case-by-case basis."

Indeed, publishers get rather hot under the collar when they are told that they are withholding permission from researchers who want to text mine their journals. Responding to an article on text mining in *The Guardian*, for instance, Taylor wrote, "To say that text mining is 'forbidden' and 'prevented' by publishers is as we have grown to expect from *The Guardian* a tendentious and limited analysis."

Apparently conceding that there may be pent-up demand for text mining, Taylor then said, "Publishers collectively fully recognise the rapidly rising demand among researchers to use text mining tools on large databases of content across several publisher platforms. But some practical measures are needed to enable that."

Researchers dispute publishers' claims that getting permission to text mine is either easy or straightforward. Whatever publishers might say, and however co-operative they may claim to be in public, they complain, in reality they are unresponsive and obstructive. And as evidence they point to the list of responses that Max Haeussler, a post-doctoral researcher at UC Santa Cruz, received from publishers when he approached them about text mining. Likewise, they point to the list of responses that Murray-Rust received.

One of the greatest problems, it seems, is that publishers often don't reply to requests from researchers, assuming that researchers know whom to contact in the first place. Publishers don't advertise this information.

Readers of the GOAL mailing list, however, might wonder if researchers are themselves sometimes a little unresponsive. Replying to a complaint about text mining that Murray-Rust had posted in May, Elsevier's director of universal access Alicia Wise said, "[W]e are happy in principle for you to mine our content … there are only some practical issues to resolve. We have successfully concluded the technical discussion, and I believe you, your colleagues, and my technical colleagues are all happy with the proposed technical mechanism."

When I asked Murray-Rust if he concurred with Wise that Elsevier had as good as agreed conditions for him to text mine its journals, he replied. "I don't want to use Elsevier's API. That means 100 APIs for me to learn — one per publisher."

The nub of the issue, it seems is that researchers resent publishers' proprietorial approach, and are thus reluctant to comply with publisher-dictated rules. "In fact, I only need a single API — a DOI resolver," Murray-Rust told me. "I may wish to systematically mine a single publisher — in which case I use a list of their DOIs, or I may want to follow links — that's exactly the same process. Yes I need an API per publisher but I and others are hacking this and it's a one-off. So a publisher API makes it worse."

And thus the impasse continues. Researchers complain that they are being blocked from text mining: publishers respond that they are happy to oblige, but must be allowed to define the terms under which it takes place, and to insist that researchers use their APIs. Researchers believe that they should be allowed to use their own tools, and their own machines, and argue that where a subscription has been paid, text mining should be viewed as an automatic right; publishers insist that text mining rights must be separately negotiated, and on a case-by-case basis.

Finally, in thinking about what has changed since I spoke to Murray-Rust in 2008, we should not forget to mention that the practice of text mining scholarly journals has gained a lot of mindshare as a result of the dogged persistence of open science advocates like Murray-Rust — and research funders like the Wellcome Trust.

Content Mining and Open Access

We said earlier that Murray-Rust was disappointed with the Open Access movement. It is therefore worth asking where exactly OA fits with text mining.

At first sight, the aims and goals of text mining and OA would appear to be exactly the same. After all, the 2001 Budapest Open Access Initiative (BOAI) stated, "By 'Open Access' to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, should be to give authors control over the integrity of their work and the right to be properly acknowledged and cited."

This would seem to imply that OA offers Murray-Rust everything he needs. As we noted earlier, however, the OA movement has not pursued reuse rights nearly as vigorously as it might have. Even today many OA publishers still do not use CC-BY licences. Indeed, OA advocate Peter Suber estimates that 88% of OA journals still do not do so.

Unsurprisingly, therefore, in 2010 only 41% of the OA content in UKPMC (the UK version of PubMed Central) was free to read and to reuse (although this is up from 30% in 2009 and 7% in 2001). More depressing for text miners, Suber estimates that the proportion of PubMed Central papers that offer reuse rights is only 18.75% — which may partly answer Bergman's question as to why so few researchers are mining the database.

It does not help that the Open Access movement frequently succumbs to bouts of acrimonious disagreement about Gratis vs. Libre OA, and a powerful constituency within the movement continues to argue that it is sufficient for researchers to be able to read papers with their eyeballs.

In May, self-styled archivangelist Stevan Harnad even argued on the GOAL mailing list that BOAI had been revised in 2008. "Yes, further re-use rights are important, and desirable, in many (not all) cases. But they are even harder to agree on and provide than Gratis OA, and we have not yet even managed to mandate that in anywhere sufficient numbers. And access itself — 'mere' access — is not just important, but essential, and urgent, for all peer-reviewed research.

For that reason, Harnad added, "it's time to stop letting the best get in the way of the better: Let's forget about Libre and Gold OA until we have managed to mandate Green Gratis OA universally."

For text miners such talk is anathema, since it implies that they should just sit on their hands until some uncertain point in the future.

However, one could in any case argue that OA is only tangentially relevant to any discussion about text mining. After all, today only 20% of papers are OA, with the remaining 80% still locked behind subscription paywalls. (And of the 20% that are OA many do not come with reuse rights). Most of the papers that text miners want to access, therefore, have been published in subscription journals.

But while that may make text mining appear to be almost exclusively a subscription issue, we could point out that those who support so-called Green OA advocate for the self-archiving of papers that have been published in subscription journals. As such, the aim of Green OA is not so different from the aim of text mining — to enable better access to content that resides behind a paywall.

But whatever the appropriate relationship ought to be between text mining and OA, we should not doubt that the former has greatly benefited from the latter in recent months. Indeed the successful protest against the recent attempt to introduce the US Research Works Act (RWA) may have done as much for text mining as it has for OA.

To remind ourselves: the RWA would have outlawed the poster child of the OA movement — the National Institutes of Health's Public Access policy. Anger at the very thought of this was sufficient to spark a boycott of Elsevier, the main supporter of the proposed legislation. (Currently over 12,000 researchers have signed the boycott).

This in turn inspired the formation last month of the Access2Research group, whose first act was to launch a petition to the US government, a petition that reads (my emphasis), "Requiring the published results of taxpayer-funded research to be posted on the Internet in human ***and machine readable*** form would provide access to patients and caregivers, students and their teachers, researchers, entrepreneurs, and other taxpayers who paid for the research. Expanding access would speed the research process and increase the return on our investment in scientific research."

In other words, the petition bolted together the notions of text mining and OA as if one and the same thing.

As it happens, it was a highly successful petition, reaching the threshold of 25,000 signatures required to ensure an official response from the Obama Administration within two weeks (by 3rd June).

Case-by-case

More significantly perhaps, as a result of the bruising that Elsevier was receiving from the boycott the publisher suddenly became much more receptive to text mining. Six days before Elsevier's ignominious withdrawal from the RWA (on 21st February), for instance, a researcher at the University of British Columbia (UBC) called Heather Piwowar was startled to receive a response to one of her tweets about text mining from Elsevier's Alicia Wise.

What followed was unprecedented: a postdoctoral researcher found herself at the centre of a negotiation over text mining between ***six*** Elsevier employees, herself, and her institutional librarian. The upshot: agreement was reached to allow Piwowar to mine Elsevier's database of journal articles. A still surprised Piwowar subsequently documented the broad details of the negotiation on her blog.

Sufficiently remarkable was the incident that it led to a series of articles in the press — including one in *The Chronicle of Higher Education*, and one in *The Guardian* — turning Piwowar into a bit of a science star, and earning her the honour of a SPARC interview — an interview that began, "In the face of boycotts and bad publicity, are publishers realising they must loosen their tight grip on usage restrictions as well?"

However it was immediately clear to Piwowar's text-mining colleagues that the method utilised by Elsevier simply would not scale. Six publishers, a librarian, and a researcher all devoting a large chunk of their time to come to a single agreement, they argued, makes no sense whatsoever.

There was also a sense that Piwowar had been slightly railroaded by Elsevier. Murray-Rust made these points at the time on Piwowar's blog. "By dealing with Elsevier you have implicitly agreed that Elsevier has the right to control what you do. That they will then generously allow you a subset of the rights that they currently deny us. If all universities follow the course of UBC we shall end up in a situation where Elsevier's walled garden philosophy controls all of us. We have a fundamental right to text-mine the literature. This agreement has given that up. I am sure it was well intentioned but that's the effect."

For all that, Piwowar attracted a lot of publicity for the text mining cause. So too have the activities of Murray-Rust, Max Haeussler and Casey Bergman (the latter two run the text2genome project); and as the aspirations of text-miners repeatedly bump up against the constraints imposed by publishers, so a growing number of incidents can be expected to attract further publicity, and consequently more mindshare, for text mining.

On 7th March, for instance, *Nature* published both an article, and an editorial on text mining highlighting the difficulties that Haeussler and Bergman had been experiencing.

The editorial concluded, "Publishers should agree that scientists who have already paid for access to research papers may text-mine content at no extra cost and publish their findings — as long as their doing so does not breach the original firewall. Publishers can have no claim on the data in articles, only on the way in which the articles have been edited and formatted."

Large gap remains

Nevertheless, a large gap remains between researchers and publishers when it comes to text mining. Researchers believe that when their institutional library pays a publisher for a collection of electronic journals it is buying the right for researchers not just to read the content with their eyeballs, but to mine it with their computers. As Murray-Rust puts it, "The right to read is the right to mine".

Nature aside perhaps, publishers appear to take the view that text mining should not be viewed as an automatic right for subscribers, and while some publishers now evidently accept that text mining will have to be countenanced, they maintain that the right to do so is separate to the right to read, and so must be negotiated on a case-by-case basis (preferably with the institutional library rather than with researchers themselves). Others remain unwilling even to contemplate it (or will only permit it on payment of an additional fee — again, on a case-by-case basis).

But there does seem to be some movement. For instance, publishers appear to have conceded that agreeing text-mining rights on a case-by-case is not very realistic. For that reason, Taylor explained in *The Guardian*, publishers are "looking into model licences, a clearing house for permissions, a collective licence to support the 'smaller' publishers, a guide for those short on 'understanding'', even a mine itself through CrossRef."

But little of this may prove acceptable to researchers, not least because they believe it is too late for publishers to start making gestures and offering concessions, and expect researchers to agree to have the conditions for access unilaterally determined. They have become too angry, and too alienated. Above all, they do not accept that publishers have the right to dictate onerous terms and conditions for accessing content to which their institution has already paid a licensing fee.

So what is the next step?

A month or so after Piwowar documented her conversations with Elsevier she posted a new note on her blog suggesting that, "The time has come for researchers to clearly state how we expect to be able to ***use*** the already-published literature."

She added, "We expect to access and process the full text of the research literature with our computer programs. We expect to disseminate aggregate statistical results as facts and context text as fair use excerpts, openly with no restrictions other than attribution. We expect these rights without further cost when papers are accessed through researcher-provided tools, and with (at most) a transparent per-api-call fee when accessed through publisher-supplied programmatic interfaces."

Declaration

Piwowar's post sparked a fertile discussion, which ended with Murray-Rust proposing that a group of like-minded people get together to draft a "Content Mining Declaration".

In short, text-mining aficionados have concluded that it is not enough to carry on advocating as they have been, and hope that governments, funders, or the OA movement, will at some point resolve matters for them; and they have concluded that they would be ill-advised to allow publishers to unilaterally determine the rules for text mining.

Rather, they believe they themselves need to "declare" what they believe to be appropriate, and acceptable. And for them to do so, they argue, is entirely reasonable given that the content in scholarly journals (including the text and data) is created by researchers in the first place and freely given to publishers — who then sell access to it back to the research community for a fee. Publicly-funded data, they believe, should be viewed as a public good, not private property.

As Murray-Rust put it on his blog last November, "We are at a critical point where unless we take action our scholarly rights will be further eroded … the science and technology of text mining is systematically restricted by scholarly publishers to the serious detriment of the utilisation of publicly funded research."

With publishers already on the back foot in the wake of the RWA debacle, the revolt over text mining has been interpreted by observers as an important new development in the OA movement.

The day after Piwowar proposed her manifesto, for instance, Claudio Aspesi, an analyst based at the sell-side research firm Sanford Bernstein, sat down and wrote an investment report on Elsevier. In doing so, he argued that "a new front" was opening up in the Open Access debate — "as academics are increasingly protesting the limitations to the usage of the information and data contained in the articles published through subscription models, and — in particular — to the practice of text mining articles."

Aspesi added, "In this instance, once again, Elsevier is the focus of much of the outrage, regardless of whether other commercial subscription publishers do or do not adopt similar restrictions. The arguments which academics are putting forward could further inflame the Open Access debate by leading critics to conclude that commercial subscription publishers, in addition to charging excessive prices for accessing research, are hindering the work of researchers as well."

But drawing up a manifesto to cater for a situation that text miners soon realised is both complex and uncertain will require a good deal of thought. And for the moment, a number of issues remain unresolved.

What, for instance, can researchers legally do with the data that emerges from their mining activities? After all, the authors of the papers will all have assigned copyright in their work to the publisher, whatever the rights and wrongs of their having done so. And as Hargreaves indicated, the legal situation on this is far from clear.

And should researchers insist that — as a non-negotiable right — they must be allowed to use their own tools to mine the data; or should they accept that access to the data will have to be mediated by publisher-controlled tools (for which there will likely be a charge)? Either way, some researchers are adamant that it is not reasonable for publishers to expect payment for use of their APIs.

Persuaded by the argument

Whatever the outcome of Murray-Rust's current initiative, governments and funders appear to have concluded that text mining is an important and valuable new research activity, and so should be facilitated. Indeed, they may even now be keener to encourage open data initiatives than open access, not least because they have been persuaded that the financial payback is potentially substantial.

When the UK Minister of State for Universities and Science David Willetts spoke to the Publishers Association in May, for instance, he asserted, "I am persuaded by the argument that we are going to see a new era of data-intensive scientific discovery."

He added, "Data mining is becoming an important part of scientific advance, with computer scientists working collaboratively with researchers and publishers to develop the necessary tools and technologies. With well over a million academic articles every year, researchers wanting to keep abreast of developments in their field are going to need analytic tools just to know where to start. There are proven benefits for humankind from text and data mining, such as the discovery of new treatments for Alzheimer's.

Consequently, Willetts said, "[W]e are considering how to advance UK capability in data mining in the light of the recommendations on intellectual property from Ian Hargreaves."

Willetts also indicated that there is a need for data mining to take place in an open environment and, like the Access2Research petition, assumed that OA and open data are but component parts of the same thing. And he stressed that he is convinced by the economic argument. "The evidence underpinning our ambition for public access is compelling," he told publishers. "For example, publicly funded and freely available information from the Human Genome Project led to greater take up of knowledge and commercialisation than from earlier protected data. To date, in fact, every dollar of federal investment in the Human Genome Project has helped generate $141 for the US economy."

Finally, Willetts reminded his audience that the UK is home to the first Open Data Institute, for which the UK government had provided £10 million in funding.

Importantly, it is not just UK politicians that have finally "got it". At the end of last year, the EC announced a new Open Data Strategy. And for the first time this will include libraries, museums and archives, with data having to be released with reuse rights — for any purpose, commercial or non-commercial, unless protected by third party copyright.

In announcing the strategy Vice-President of the European Commission responsible for the Digital Agenda Neelie Kroes said, "We are sending a strong signal to administrations today. Your data is worth more if you give it away. So start releasing it now: use this framework to join the other smart leaders who are already gaining from embracing open data. Taxpayers have already paid for this information, the least we can do is give it back to those who want to use it in new ways that help people and create jobs and growth."

Once again, we should note that both the above initiatives are primarily focused on data generated by governments and public bodies, not scientists. But the more the spotlight on open data (of whatever sort) is turned up, and the more governments conclude that open data and data mining promise significant economic benefits, the more likely it is that Murray-Rust and his text-mining colleagues will get what they want.

Statement of fundamental rights

The extent to which open data and text mining are infiltrating the OA debate was evident at the Publishing and the Ecology of European Research (PEER) Project Conference held in Brussels at the end of May.

The conference was organised to report on developments in OA, but Commissioner Kroes moved seamlessly from the topic of access to journals, to open data (and by implication to text mining). "[W]e should not limit ourselves to journal articles and the like," she said. "Open access to research data, too, would open a new field of opportunity. Meaning you can re-analyse experiments; boost the impact of research; and provide a precious fuel for new collaborations and new knowledge-based industries. Those open data benefits, direct and indirect, can't be ignored."

Commissioner Kroes concluded, "[W]hen research is funded by the EU, we will require open access to the results. Whether by 'green' or 'gold' routes. And we're working to enlarge those measures to include scientific data as well."

But after sitting in the wilderness all these years, Murray-Rust is determined not to leave matters to chance, or the unsteady hands of politicians. It is essential, he says, to articulate and publish the precise needs and requirements of text miners before governmental agendas are set in stone, or dissolve under the pressure of "events". As such, he says, the proposed Content Mining Declaration should be viewed as an equivalent to the BOAI, but focused specifically on content mining. Above all, he adds, it will be a statement of principles, not an overture to a discussion with publishers.

Murray-Rust's fears are twofold. First, conscious of the growing pressure to permit text mining, publishers will surely soon concede the point, but in so doing seek to dictate the terms on which it can take place. As noted earlier, publishers are determined to maintain control, not least because by doing so they can hope to earn additional revenues from licensing APIs and other intermediary software services.

Second, with governments so focused on the economic aspects, they may be tempted to reach a settlement with publishers that prioritises the needs of publishers over the needs of the research community. In his speech to the Publishers Association, for instance, Willetts said, "Provided we all recognise that open access is on its way, we can then work together to ensure that the valuable functions you carry out continue to be properly funded — and that the publishing industry remains a significant contributor to the UK economy … We recognise the value which publishers add … It would be deeply irresponsible to get rid of one business model and not put anything in its place."

The latter danger seems likely if one considers what is currently happening in the OA space: While publishers have conceded the need to embrace OA, they want to provide it exclusively by means of Gold OA, not Green OA — because Gold OA holds out the promise of enabling them to lock their current revenues into the new publishing model, revenues that many believe to be excessive. The problem, argue critics, is that this approach will prevent the research community from resolving the underlying affordability problem that motivated many to advocate for OA in the first place.

Murray-Rust therefore believes that it is essential for the research community to set the terms, not wait for publishers to do so. Writing about the proposed declaration on his blog, he said, "This isn't a negotiated position. It's not a summary of current practice. It's a statement of a fundamental right that we must fight for."

Indeed, he feels so passionately about "the right to mine" that he likens it to basic political freedoms like free speech. "In the 20th Century the people asserted their right to roam. We are

asserting the people's right to mine. This is a simple political statement — like 'everyone has a right to a fair trial'. Because the publishers — like the 19th C landowners dispute this right we have to fight for it. The UK has had a series of fights for rights including freedom of speech, trial by jury, freedom from slavery, etc. Sometimes people went to jail, sometimes they died for these. But we must fight."

Murray-Rust is clearly a determined man. He is also very keen that the wording of the Declaration is fit for purpose. To that end he has released the draft text, and is seeking input from others. Those wishing to comment can view the draft text here, or read extracts of it on Murray-Rust's blog here.

Q&A with Peter Murray-Rust

RP: When I interviewed you in 2008, your focus was on what you called open data. Today it is on text/data/content mining. What journey took you from open data to content mining?

P M-R: I want to do data-driven science, particularly in physical science. I believe there are patterns hidden in existing data. I first started this in 1974 with crystallography, as the data was somewhat available — although I had to copy a great deal out of paper journals.

In 1978, I was able to use the Cambridge Database to do this and I set off a revolution in how crystallography could be re-used as a result.

I then decided I wanted to do this in other subjects like computational chemistry and spectroscopy, and I hoped that authors would publish machine-readable data files to enable this. They generally have not, however, even in the last 35 years.

By creating the idea of "open data" (and surprisingly I was the first person to make use of the term for mainstream use, about six years ago) I hoped to spark a wave of interest. This has happened (not really due to me) but not in many physical sciences.

So the only way forward at the moment is content-mining. It's not fun. I have spent five solid — very solid — weeks of 14/7 code writing to develop tools for this.

RP: The point then is that as researchers are still not providing this data in machine-readable form themselves, the only way of getting at it is by extracting it yourself from the text. But how would you describe contenting mining? And how does it differ from text mining and data mining?

P M-R: There is a hierarchy. Content mining includes text mining but also others types of mining such as images, tables, graphs, audio, and video. The problem with using the word "text" is that it might imply that anything other than text should be protected by the "content provider". Others and I can extract factual information from a wide range of material.

But let's be clear, today the only real problem remaining is the publishers. The community now understands mining. Therefore, there is increased pressure and I think we shall get change or a seismic fracture. I think the latter would be more beneficial so long as we do not let yet-another-robber-baron appropriate it.

RP: What is the purpose of the Content Mining Declaration you are working on?

P M-R: The aim is to do the following:

To assert the need and value of content mining.

To assert the rights and responsibilities of content miners.

To assert that this should be an open process and to define the scope of the openness.

To focus the community on best practices and avoid ad hoc approaches that later turn out to be problematic.

To emphasise that technical and permission barriers, however small, are serious impediments to mining.

Our base claim is that the right to read is the right to mine.

Note that the motivation for this springs from the value and need to content mine in scholarly publishing, but it applies to any field where people and their machines have legitimate access to content and where it is legitimate to make the mined content open. So while the primary motivation is factual data in scientific publications, it could extend to, say, government reports.

RP: How is content mining related to Open Access? Is it related?

P M-R: "Open Access" is used in a wide variety of ways and is usually not precisely defined. Where "Open Access" means that material is publicly visible the Declaration will stress a need for a clear statement (for both humans and machines) as to whether the content is minable without further permission. In general this is only provided by a machine-readable licence such as CC-BY or CC0.

Unless defined in this way, "Open Access" does not assert the right for the content to be mined. Implementation of the Declaration by the content provider will make this clear.

RP: Why do you think a Content Mining Declaration is necessary, and why now?

P M-R: There is currently enormous interest in this area and it has been estimated to open up a $100+ Billion market and provide enormously better use of science funding.

Because it has been forbidden so far, it is difficult to predict all the new value, but these include indexing science on facts, validating against other scientific facts, and the discovery of new science by aggregating facts.

RP: Was there a particular incident or conversation that convinced you the Declaration was necessary?

P M-R: Not a single point. I have been building Content Mining tools for 10 years but in the last year enough things have come together to suggest it will take off.

It is clear that practice at all levels is fragmented and if 100 universities, 10,000 scientists and 100 publishers try to make the rules on the fly we shall have chaos and very bad value for our effort.

RP: Graham Taylor of the Publishers Association maintains that when researchers ask publishers for permission to text mine publishers generally grant it "willingly and easily" for non-commercial purposes. He also says that publishers are currently looking into model licences and a clearinghouse for permissions. You do not think this is sufficient. Why?

P M-R: Because it is completely opposite to my experience and I think many others. It has taken one researcher (Piwowar) weeks of time with her library and Elsevier to come up with an agreement that doesn't scale. A simple example of the kind of obstacles we face in getting agreement is that no publisher makes it clear who should be contacted.

Publishers (in the main) are commercial companies and models will be oriented towards their needs, not researchers and humanity in general. In any case, it is about time the "customers" said what they wanted, not waited for the publishers to create something.

RP: In May, Elsevier's Alicia Wise posted a message on the GOAL mailing implying that she had as good as sorted out your text mining needs at Cambridge. Is that your view? If not, what are the problems with what she is offering you?

P M-R: I don't want to use Elsevier's API. That means 100 APIs for me to learn — one per publisher.

In fact, I only need a single API — a DOI resolver. I may wish to systematically mine a single publisher — in which case I use a list of their DOIs, or I may want to follow links — that's exactly the same process. Yes I need an API per publisher but I and others are hacking this and it's a one-off. So a publisher API makes it worse.

The only conceivable argument for publishers to insist on the use of APIs is server load. Elsevier publishes 250,000 papers each year, has an archive of 7 million publications, and has 240 million downloads. That's a soluble problem.

RP: Is it not likely that the UK government will in any case resolve the problem for you? The Hargreaves Report proposed an exception to copyright law to support text mining and analytics, including provision to override any attempt to set it aside in the words of a contract. I understand the government has accepted this recommendation.

P M-R: Firstly we actually have to get Hargreaves implemented. My sources tell me this is likely but in the world of politics there may be "events" (Harold Macmillan). Secondly it will apply to the UK only.

RP: What is the process you are going through to produce the Declaration?

P M-R: We have put together a small group of people intimately involved in content mining and open issues to create a draft. I have posted extracts on my blog.

RP: Are you looking for input from others? If so, what kind of input, and from whom?

P M-R: Yes. We are exposing the full draft for public comment. The main input we are looking for are comments from those whose interests are aligned with ours and who may add clarification, additional material or political weight. But I stress, it is not a community agreement, it's a declaration.