The social, political and legal aspects of text and data mining (TDM).

BROCK, M., MURRAY-RUST, P. and OPPENHEIM, C.

2014

D-Lib Magazine

November/December 2014
Volume 20, Number 11/12

The Social, Political and Legal Aspects of Text and Data Mining (TDM)

Michelle Brook
The Content Mine
michelle@contentmine.org

Peter Murray-Rust
University of Cambridge
pm286@cam.ac.uk

Charles Oppenheim
City, Northampton and Robert Gordon Universities
c.oppenheim@btinternet.com

doi:10.1045/november14-brook

Abstract

The ideas of textual or data mining (TDM) and subsequent analysis go back hundreds if not thousands of years. Originally carried out manually, textual and data analysis has long been a tool which has enabled new insights to be drawn from text corpora. However, for the potential benefits of TDM to be unlocked, a number of non-technological barriers need to be overcome. These include legal uncertainty resulting from complicated copyright, database rights and licensing, the fact that some publishers are not currently embracing the opportunities TDM offers the academic community, and a lack of awareness of TDM among many academics, alongside a skills gap.

1. Introduction

The ideas of textual or data analysis go back hundreds if not thousands of years. Originally carried out manually textual and data analysis has long been a tool which has enabled new insights to be drawn from text and data corpora.

The development of the computer enabled the automation of the processes involved, including information retrieval, searching, indexing, reading and analysis of literature that exists in a digital format. These automated processes, often referred to as text and data mining (TDM), have many everyday uses, including the underlying technology behind Internet search engines and targeted advertising, a variety of ways to access library collections, and sentiment. There are many potential uses for TDM within academia, including within biomedical, biological and chemical research where genes, species names, chemical compounds, and statistical significances can all be extracted. Clark (2012) provides a useful introduction to the topic.

Although the computer-based techniques date back to the early 1980s, unsurprisingly, the development of TDM as a field has relied heavily upon improvements in computational hardware, including lowered memory and CPU costs, development of high speed networks, as well as innovations in software, including natural language processing techniques, and optical character recognition.

Any content undergoing analysis has to exist in what is termed a 'machine readable format', one of several formats known as XML or JSON that allow a computer to read and understand the information contained. PDFs, often the *lingua franca* of academic journals, while able to present standardized formats for humans to read, are not machine readable, and as such for TDM purposes, this work must be transferred into a different digital form (Hobbs, *et al*., 1982). That form is often custom and specific to the research question being asked and the most appropriate tools to answer that question.

There is, unsurprisingly, a vast literature describing research into, and applications of, TDM, as well as the technical challenges. However, to date relatively little has been written on the political, social and legal barriers involved. The primary legal issue associated with TDM relates to copyright, database rights and licensing4, to which there are political solutions. The lack of awareness, and relative technological gap between many TDM tools and the skills of many academics also acts as a barrier.

Despite the existence of technological and legal barriers, there has been a growth in interest in TDM over recent years; including the creation of the first publicly funded text-mining centre, the National Centre for Text Mining in Manchester in 2004, TDM being the subject of various reports, including a 2012 JISC report (McDonald and Kelly, 2012) and the development of Google Ngram. This growth in interest is not simply a reflection of the improvements in techniques and technologies. There has been recognition that TDM technologies provide broad economic and societal opportunities, 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." (McDonald and Kelly, 2012)

However, for the potential benefits of TDM to be unlocked, a number of non- technological barriers need to be overcome. These include legal uncertainty resulting from complicated copyright, database rights and licensing, the fact that some publishers are not currently embracing the opportunities TDM offers the academic community, and a lack of awareness and technological skills among many academics with respect to TDM. These are not the only barriers to the uptake of TDM; however, topics such as ethical and privacy issues for medical data, the fact that data quality can be poor, e.g., data held in different systems, missing, corrupted, non-standardised data; the fact that sometimes patterns emerge which are in fact due to random fluctuations; and the fact that setting up a sophisticated TDM facility involves substantial investment in IT systems, building databanks and recruiting expertise are not considered further in this paper.

2. Copyright, database right, licences and TDM

The primary legal issues facing anyone wishing to undertake TDM are copyright law, database law and contract law (the licensing of work)[1]. While the issues raised below are relevant worldwide, the comments regarding legal standing below are specific to UK law, and readers should not assume they are applicable to other jurisdictions. There will however be similarities in these other jurisdictions.

Copyright law, many claim, states that anyone creating and recording in any manner a new creative work (i.e., one not copied from elsewhere) is automatically granted copyright protection over that work. This protection gives the owner the right to authorise, or to refuse to authorise, certain 'restricted acts' by any third party; these acts include copying, adaptation of the work, redissemination (for instance publishing it on the Web), all, or a "substantial part" of the original copyright work, as well as translation of the work into other languages.

In practice, copyright law is more complex than stated above. Simple facts (which includes much data) are not subject to copyright, and nor are short sentences. What is classified as either a fact or a short sentence remains ambiguous and controversial, and is often only resolved by means of a court case. A recent European Court case[2] confirmed that a sentence of 11 words was capable of being protected by copyright; though this does not imply that all similar sentences will be protected by copyright.

As the act of copying is considered to infringe the Intellectual Property Rights of copyright owners, this leads to some legal ambiguity around the process of TDM. Some processes of TDM, from a PDF, may require the production of a temporary interstitial file. Whether those files are "legal" is focus of debate between those who support or oppose wider TDM use. Is an interstitial file the equivalent of a temporary file from Adobe Acrobat Reader? Or is it a derived work and therefore a breach of copyright?

These questions are significant. As the majority of the academic literature currently exists under copyright, without licenses relinquishing rights, we can assume much of the literature that would be used for TDM would be under copyright. While there are some so-called exceptions to copyright[3], which permit third parties to copy, or carry out restricted acts, without having to pay or having to request permission and although such exceptions in some countries such as the United States, are considered sufficient to allow for content mining, researchers may be reluctant to test the law which is not well defined for these activities. In the UK, until a recent change to copyright law, there were numerous copyright exceptions, however the wording of these were either ambiguous or arguably did not apply to TDM.

This results in a legal barrier, which can prevent risk-averse researchers (or their institutions) from relying upon copyright exceptions to carry out their research, and instead approaching publishers (as most of the works they wanted to use were indeed owned by publishers) for permission to TDM for a particular project for a particular pre-defined question. Sometimes this permission would be readily granted at no charge, but many publishers wanted unreasonably high fees and/or placed such restrictions on what could be done with their materials and/or took an extremely long time to decide how to respond, so that in practice the permission was not obtainable. There are some notable exceptions, such as Springer and The Royal Society, that have taken a generous approach to granting permissions to researchers to undertake TDM, but alas they are still a minority. In any case, TDM researchers have to approach multiple publishers, each of whom have different attitudes, conditions, and speed of response to such requests. This is very costly to a researcher, and would have significant impact upon the take-up of TDM as a practice, or (so anecdotes tell us) inhibit academics from sharing the outputs of their research using TDM, therefore limiting the potential benefits this technology enables.

3. Recent changes to UK law

In 2010, the UK Prime Minister commissioned a report to consider whether the UK's intellectual property framework was up to the task of supporting innovation and growth (the Hargreaves Report) (Hargreaves, 2011). One of the key recommendations (Chapter 5.26) was that the UK government should introduce a copyright exception to allow the use of analytics for non-commercial use. It is not clear where Hargreaves developed this idea from, but it must in part have been based on evidence submitted by the UK library and research communities, e.g. (LACA, 2011). In December 2012, the UK Intellectual Property Office carried out an impact assessment for the introduction of such an "exception for copying of works for use by text and data analytics". For that purpose, it defined "text and data and data analytics methods" as methods to: "extract data from existing electronic information, to establish new facts and relationships, building new scientific findings from prior research. These new methods involve copying of prior works as part of the process to extract data" (UK IPO, 2012). It did not attempt to put a figure on the likely economic impact of such a change to the law, but noted that in the longer term, the change would lead to "social innovation and longer term scope for major economic gains."

On the 1 June 2014, the UK Parliament passed legislation to change the law on TDM, stating that:

> "The making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that —
>
> > (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and
> >
> > (b) the copy is accompanied by a sufficient acknowledgement (unless this would be impossible for reasons of practicality or otherwise)."

A critical rider states that:

> "To the extent that a term of a contract purports to prevent or restrict the making of a copy which, by virtue of this section, would not infringe copyright, that term is unenforceable."

The importance of this change to the law cannot be over-emphasised. Firstly, at a stroke, it puts to an end the arguments about whether TDM might or might not have been permitted under previous UK copyright exceptions. Secondly, it makes efforts by publishers to impose their own terms on UK- based researchers a waste of everyone's time and money. Thirdly, it offers a role model for other countries for legislation wording to help promote these useful techniques. Finally, at least temporarily, it offers TDM researchers in other countries a way to by-pass licensing restrictions and copyright infringement worries by working collaboratively with UK-based researchers who have lawful access to the relevant raw materials, letting the UK-based TDM researcher undertake the work, knowing that there is no legal risk involved.

It is clear that the numerous constraints have, for UK researchers, been removed at a stroke. This is not to say that the situation is perfect. There is still confusion about how widely the term "non-commercial research" might be interpreted. Clearly a researcher employed by a pharmaceutical company cannot claim to be doing "non-commercial" research, but what about a researcher in academia who has received funding from a pharmaceutical company, or has been seconded to a commercial company to undertake TDM research? Or an independent researcher who plans to publish their work on a blog, which gains income from advertising. These questions will only be answered by Court cases — something that rights-owners and TDM researchers will each be reluctant to engage in, and until these questions are resolved, there are many researchers who will remain hesitant to fully embrace these tools.

Database rights are similar to, but more limited than, those for copyright, and applies to certain collections of data (data including many entities, such as words, numbers, or images, though probably not to moving images or sound recordings). The protection of rights in databases is complex, and a simplified summary only is given here. A database can enjoy copyright (and be subject to the points above about copyright law) if the selection and arrangement of the materials in the database has involved intellectual effort. In other words, a collection of data can be protected by copyright if there is sufficient creativity involved in the presentation or arrangement of the set, although the underlying individual data points remain free of any protection.

Thus, a totally comprehensive database of records, where no selection has been undertaken, cannot enjoy copyright. But if there has been intellectual effort in making the selection, then copyright applies. In addition, and quite separately, a database can enjoy database right if effort of any kind has been expended in obtaining, verifying and presenting the data.

A database, therefore, might enjoy copyright and/or database right, or it might have no rights at all associated with it, dependent on how the database was created in the first place. As with copyright, the law here is often unclear and acts as a chilling effect and barrier to many.

Licences/contracts are important features of the TDM legal landscape. Owners of copyright and/or database rights will often grant permission to third parties to read, download, redisseminate or otherwise exploit their materials, typically for a fee (which can be substantial in the case of large scholarly publishers owning a wide range of resources). Such licences will inevitably impose terms and conditions — limitations — on what the licensee can do with the licensed materials. Increasingly, we are seeing an uptake in Creative Commons licences, which enable authors grant blanket permissions to third parties to use their work in pre-defined manners. However even these relatively liberal licenses impose some terms and conditions on the user, such as having to acknowledge the source[4]. Many non-Creative Commons licences do not allow TDM (although these licences are of course over-ridden by copyright exceptions within a specific country such as the UK, or Japan, which has a similar exception to the UK for TDM).

While some public funders, such as the UK Research Councils now require the use of Creative Commons (CC) licences for research outputs produced by their funds, many research funders around the world have not yet done the same. The use of these licences can be very useful for academics who wish to engage in TDM in those countries without appropriate copyright exceptions. Machine-readability of licences is crucial for TDM, although many publishers have yet to exploit fully the opportunities that the machine-readable CC licenses offer. Furthermore, licence proliferation should be resisted, especially in instances where the new licences produced are not interoperable with those already existing. This is especially important in an international context, given ambiguities in interpreting copyright law within different jurisdictions. With Creative Commons licences, the *de facto* standard in Open Access licensing, and increasingly within open data, one has question why publishers feel the requirement to try to impose control in this fashion, and why many of these licences are framed as 'Open Access', despite not conforming to the BOAI definition of Open Access.

A recent example of licence proliferation can be seen from The International Association of Scientific, Technical and Medical Publishers who recently promoted their own set of licences (STM, 2014). An introduction to these licences states, without any supporting evidence, "the complexity of stakeholder interests and commitments means that no one solution that is imposed externally can be effective, whether by way of national exceptions or otherwise: solutions must therefore be worked out by multi-stakeholder dialogues which respect stakeholders' interests, are conducted in a transparent manner and aim at models that are simple and scalable." Such publisher-driven TDM licences have been the subject of fierce, and justified criticism by the library (LIBER News, 2014) and researcher communities because the proposed licences are too limiting and impose unfair (and in the case of the UK, unenforceable) constraints on researchers' freedom to exploit TDM.

4. What can politicians and policy makers do?

In many countries, the law relating to TDM is ambiguous, and this will inhibit researchers from undertaking it. Copyright and database rights should evolve to reflect practices and technologies of the time, not inhibit the development of new research. Where ambiguities lie, these should be removed. For instance, under current EU law, countries in the European Union are able to introduce exceptions for non-commercial TDM research, but to date it is only the United Kingdom which has taken advantage of this. There has been no change to date at the EC level, although they are considering an EU-wide exception for TDM, and the Republic of Ireland is also considering such a change to its law. Outside of Europe, the authors have only heard of one or two other Far East countries that have introduced such exceptions. This is a significant barrier to TDM worldwide.

While it is clear that the UK is leading the way here, there is nonetheless, ambiguity remains regarding what is, or is not, "commercial research", and it may well require some court cases to clarify the situation, although the ambiguity is likely to have a chilling effect until these questions are resolved. Research funders can encourage the use of Creative Commons licences for research outputs resulting from their funding, and are in a position to engage with their fundees about the problems facing licence proliferation.

5. Publishers are not embracing opportunities of TDM

It is clear that some publishers are very reluctant to give up the control they currently have over how TDM is carried out on materials where they own the copyright. A typical statement comes from Richard Mollett, the head of the UK Publishers Association: "we support content mining. It can only work well if we are involved in the process and managing the access." [5] The guidance to the legislation introduced in response to the Hargreaves legislation specifically states that publishers can **only** apply 'technological measures on networks that are required in order to maintain security or stability but won't be able to enforce contract terms that seek to prevent or unreasonably restrict text and data mining", and it will be interesting to see if publishers recognise that, e.g., **requiring** a TDM researcher to use only its "approved" API is not lawful.

6. How can publishers help TDM researchers?

Publishers can do much to help TDM researchers. Given the recent changes to UK law, the fact that other countries now either have similar laws, or are proposing to introduce them (and indeed, there is now pressure for an EU Directive offering such an exception throughout the European Union), publishers can help TDM research by dropping efforts to "control" the process, but rather by pragmatically adopting the following policies:

1. Offer all researchers world-wide the same freedom as is now available to UK researchers to undertake TDM for non-commercial research purposes, so long as the user has lawful access to the original materials, i.e., follow the lead of Springer and The Royal Society.[6]
2. Earn goodwill amongst the TDM research community by offering user- friendly APIs (without, of course, REQUIRING a researcher to use them), free advice, and discussion fora for the exchange of experience and ideas in the theory and practice of TDM — in other words, by adopting a policy of welcoming and encouraging the change rather than appearing to resist it.
3. Develop clear agreed statements as to what types of research they agree is "non-commercial" and which is "commercial". This could usefully be done by a joint working party of publishers, TDM researchers and JISC. There is precedent for this type of collaboration — the Publisher and Library Solutions (PALS) developed many useful agreed statements over the years.

It is no secret that commercial scholarly publishers do not enjoy a good reputation with scholarly researchers at the moment. This has been caused by cynical policies on Open Access (with publishers sometimes using the term when what is on offer is not OA at all), and so-called "double-dipping" charging for gold and subscription-based journals. The recent change in UK law gives publishers a golden opportunity to improve their reputation. But will they grasp it?

7. Awareness among academics and a technological gap

It appears to the authors that there is a lack of awareness among many academics regarding TDM technologies, what these technologies may enable, and how to use the technologies. Current TDM researchers are very technologically adept and work will need to be done to develop the existing tools to be easier for utilisation by those with less expertise. While The Content Mine is beginning to redress some of this, developing training workshops on TDM, and other organisations such as Software Carpentry are running workshops to help academics become more technically confident, much more needs to be done. Some of this will need to come from the wider TDM community, who need to help close the gaps in knowledge, ability and awareness, but funders and institutions also have a responsibility to help ensure academics are trained on such skills and technologies; by running workshops, incorporating training sessions into PhD skill sessions, and by production of information for academics to read and understand.

8. Conclusion

The main barriers against the uptake of TDM are not technical, but primarily a lack of awareness among academics, and a skills gap. They relate to legal issues around copyright and database rights, and to some policy choices of restrictions being implemented by publishers on (for instance) access to APIs. These problems are all soluble, but require non-technical solutions.

Within the UK there are now copyright exceptions for TDM. Other countries should look to implement similar, perhaps using wording from UK legislation as a template, making it clear that TDM is permissible within their jurisdiction. Publishers should work with the TDM academic community to develop agreed statements as to what types of research they agree is "non- commercial" and which is "commercial", and prevent any possible chilling effect from ambiguity around these terms. Funders and institutions should be exploring how to teach TDM techniques to interested academics, including the possibility of opportunities to learn about it within PhD programmes. Within the UK, we have got about as far as politics can take us. What we now need to do is to show the value of TDM, to encourage other countries to develop similar copyright exceptions, and to show other academics that TDM is worth their time.

Notes

[1] In addition there are there are some related issues such as privacy and data protection that we do not explore here further

[2] See Case C-5/08, Infopaq Int'l A/S v. Danske Dagblades Forening [2009] ECR I-6569.

[3] These are the often so-called 'fair use' exceptions, or 'fair dealing' within the UK.

[4] See Creative Commons, About the licenses.

[5] See R. Mollett, Statement to Houses of Parliament.

[6] See details of the Royal Society and Springer TDM Policies.

References

[1] J. Clark, *Text mining and scholarly publishing*, 2012.

[2] J. R. Hobbs, D. E. Walker and R. A. Amsler, Natural language access to structured text, *Proceedings of COLING '82, the 9th Conference on Computational Linguistics*, 1982, 1, 127-132. http://doi.org/10.3115/991813.991833

[3] D. McDonald and U. Kelly, *The value and benefits of text mining*, 2012.

[4] I. Hargreaves, *Digital Opportunity*, 2011.

[5] LACA Response to Independent Review on Intellectual Property and Growth, 2011.

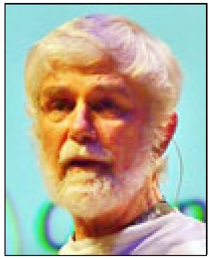
[6] UK Intellectual Property Office, *Economic Impact of Recommendations*, 2012.

[7] STM, Text and data mining: STM statement and sample licence, 2014.

[8] LIBER News, 2014.

About the Authors

Michelle Brook is a former Manager at ContentMine.



Peter Murray-Rust is Reader in Molecular Informatics at the University of Cambridge, and Senior Research Fellow of Churchill College, Cambridge. He was educated at Bootham School and Balliol College, Oxford and obtained a Doctor of Philosophy. His research interests have involved the automated analysis of data in scientific publications, creation of virtual communities and the Semantic Web. In 2002, Dr. Murray-Rust and his colleagues proposed an electronic repository for unpublished chemical data called the World Wide Molecular Matrix (WWMM). In 2014 he was awarded a Fellowship by the Shuttleworth Foundation to develop the automated mining of science from the literature. In addition to his work in chemistry, Murray-Rust is also known for his support of open access and open data.

Charles Oppenheim was Professor of Information Science and Head of the Department of Information Science at Loughborough University until retiring in 2009. He is currently a Visiting Professor at Robert Gordon University, Aberdeen, at the University of Northampton, and at Cass Business School, part of The City University, London. Previously he held posts in other academic institutions, and for twelve years worked in the electronic publishing industry. He can be followed on Twitter: @CharlesOppenh.