Supporting Document T

Text Mining and Data Analytics in *Call for Evidence* Responses

In Chapter 5 the Review makes recommendations to enable research. In the long term changes to the copyright framework need to be made at EU level, but it will be necessary to act at UK level in the interim to provide protection to those who seek to utilise text mining (or text data mining) to promote scientific research and the furthering of public knowledge.

Text mining is the process of deriving information from machine-read material. It works by copying large quantities of material, extracting the data, and recombining it to identify patterns. Copying a substantial part of a copyright work is an act restricted by copyright, and so for legal reasons it requires permission from the rights holder. A number of major organisations state that research in the UK is being blocked by the need to gain copyright permissions for access to archives. Analogue era copyright rules should not block progress towards significant scientific and cultural advancements.

There is an argument that copyright was never intended to prevent the kind of research function provided by text and data mining as it is intended to protect expressions rather than facts. Copyright law was established long before this technology was developed, and the system has not been flexible enough to provide an effective accommodation. There are strong arguments for amending the legal framework to provide a mechanism to allow for text and data mining to be undertaken without requiring permission from rights holders because the technique does not seek to make use of the expressions intended for protection.

Copyright may not be the only obstacle as access also needs to be gained to published journals. This access may be frustrated by provisions in the standard terms of contracts between publishers and institutions.

A selection of comments taken from the 247 submissions which related to copyright where text mining was mentioned are at the end of the document. The responses to the *Call for Evidence* presented in this document are intended to be illustrative of views represented in these submissions rather than an exhaustive inventory.

The Value of Text Mining

A range of case studies demonstrating the uses for text and data mining are provided in C*all for Evidence* submissions, citing examples of increasing efficiencies and increasing the speed of biomedical discovery. Research utilising text mining has facilitated the “creation of hypotheses regarding the roles of four genes never previously characterised as involved in craniofacial development a significant breakthrough with far reaching implications.”[1]

The value to scientific research is that text and data mining uses existing knowledge to discover new hidden relationships (for example, in cases where A and C have no direct relationship but are connected via shared B intermediaries) which can then be validated through experiments to test these hypotheses.

“Marc Weeber and colleagues used automated text mining tools to infer that the drug thalidomide could treat several diseases it had not been associated with before. Thalidomide was taken off the market 40 years ago, but is still the subject of research because it seems to benefit leprosy patients via their immune systems. Weeber and Grietje Molema, an immunologist, used text mining tools to search the literature for papers on thalidomide and then pick out those containing concepts related to immunology. One concept, concerning thalidomide’s ability to inhibit Interleukin-12 (IL-12), a chemical involved in the launch of an immune response, struck Molema as particularly interesting.

A second automated search for diseases that improve when the action of IL-12 is blocked revealed several not previously inked with thalidomide, including chronic hepatitis, myasthenia gravis and a type of gastritis. ‘Type in thalidomide and you get 2–3000 hits. Type in disease and you get 40,000 hits. With automated text mining tools we only had to read 100–200 abstracts and 20 or 30 full papers. We’ve created hypotheses for others to follow up,’ says Weeber.

Weeber et al. **Journal of the American Medical Informatics Association**. 2003 10 252–259” *British Library submission*

Leveraging text mining to improve human curation

“**Background**: The Comparative Toxicogenomics Database (CTD) is a publicly available resource that promotes understanding about the etiology of environmental diseases. It provides manually curated chemical-gene/protein interactions and chemical- and genedisease relationships from the peer-reviewed, published literature. The goals of the research reported here were to establish a baseline analysis of current CTD curation, develop a text-mining prototype from readily available open source components, and evaluate its potential value in augmenting curation efficiency and increasing data coverage.

Results: Prototype text-mining applications were developed and evaluated using a CTD data set consisting of manually curated molecular interactions and relationships from 1,600 documents. Preliminary results indicated that the prototype found 80% of the gene, chemical, and disease terms appearing in curated interactions. These terms were used to re-rank documents for curation, resulting in increases in mean average precision (63% for the baseline vs. 73% for a rule-based re-ranking), and in the correlation coefficient of rank vs. number of curatable interactions per document (baseline 0.14 vs. 0.38 for the rulebased re-ranking).

Conclusion: This text-mining project is unique in its integration of existing tools into a single workflow with direct application to CTD. We performed a baseline assessment of the inter-curator consistency and coverage in CTD, which allowed us to measure the potential of these integrated tools to improve prioritization of journal articles for manual curation. Our study presents a feasible and cost-effective approach for developing a text mining solution to enhance manual curation throughput and efficiency.

Thomas C Wiegers, Allan Peter Davis, K Bretonnel Cohen, Lynette Hirschman and Carolyn J Mattingly **Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database (CTD).** BMC Bioinformatics 2009, 10:326doi:10.1186/1471-2105-10-326" *British Library submission*

Discovering a New Link between Genes and Cranofacial Development

"This work tackled the problem of dealing with large amounts of experimental data (high throughput data), by combining text mining over the scientific literature, reasoning from ontologies, and networks constructed from experimental data, resulting in: "creation of hypotheses regarding the roles of four genes never previously characterized as involved in craniofacial development; each of these hypotheses was validated by further experimental work."

These hypotheses were experimentally validated by in situ hybridization and may have clinical consequences related to cleft lip and palate.

Leach SM, Tipney H, Feng W, Baumgartner WA Jr., Kasliwal P, et al. 2009 **Biomedical Discovery Acceleration, with Applications to Craniofacial Development.** PloS Computational Biology 5(3): e1000215. doi:10.1371/journal.pcbi.1000215" *British Library submission*

Discovering a New Link between Genes and Osteoparosis

"We created a text mining tool that analyzes the PubMed literature database and integrates the available genomic information to provide a detailed mapping of the genes and their interrelationships within a particular network such as osteoporosis. The results obtained from our text mining program show that existing genomic data within the PubMed database can effectively be used to predict potentially novel target genes for osteoporosis research that have not previously been reported in the literature.

Varun K. Gajendran, Jia-Ren Lin, David P. Fyhrie, **An application of bioinformatics and text mining to the discovery of novel genes related to bone biology,** Bone, Volume 40, Issue 5, May 2007, Pages 1378-1388, ISSN 8756-3282, DOI: 10.1016/j.bone.2006.12.067. (http://www.sciencedirect.com/science/article/B6T4Y-4MVVSS1 1/2/df681f901acd33d5f3eceedb36fe441e)" *British Library submission*

Uses for text and data mining have been identified outside the scope of pure scientific discovery. These include summarising legal judgements to "produce patterns that disclose habits and minds of judges and legislators that would have otherwise gone unnoticed"[2] and they are also usable in a business context. One example cited in the British Library submission to the Review is the Dow Chemical's merger with Union Carbide Corporation.[3]

"In 2001, Dow Chemicals merged with Union Carbide Corporation (UCC), requiring a massive integration of over 35,000 of UCC's reports into Dow's document management system. Dow chose ClearForest32[1], a leading developer of text-driven business solutions, to help integrate the document collection. Using technology they had developed, ClearForest indexed the documents and identified chemical substances, products, companies, and people. This allowed Dow to add more than 80 years' worth of UCC's research to their information management system and approximately 100,000 new chemical substances to their registry. When the project was complete, it was estimated that Dow spent almost $3 million less than what they would have if they had used their own existing methods for indexing documents. Dow also reduced the time spent sorting documents by 50% and reduced data errors by 10-15%.

Fan, W., Wallace, L., Rich, S., Zhang, Z. (2006) **Tapping the power of text mining.** Communications of the ACM 4(9): 76-82. See also: http://www.computerworld.com/s/article/85113/ClearForest_Scaling_Dow_s_Paper_Mountain" *British Library submission*

Supporting the education evidence portal via text mining – Increasing Efficiencies

"A project was undertaken in collaboration with James Thomas of the Institute of Education's Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI), to add text mining functionality to the UK Education Evidence Portal (eep) which is used by several government departments as well as researchers. A major activity there is systematic reviewing of the type you yourself are undertaking. The conclusions were reported in Philosophical Transactions of the Royal Society:

"Text mining services have been used to enhance search and discovery options for the UK eep. Combinations of metadata enhancement, improved browsing and navigation, alongside alternative views of resources, have all strengthened the overall proposition of the portal. Particular features include the automatic classification of lengthy documents and reports (as opposed to only abstracts) according to a custom-built, domain-specific taxonomy, automatic grouping of documents into clusters that are generated on demand according to the contents of the retrieved documents, and automatic identification of key terms within documents, which facilitates quick scanning of documents, as well as allowing closely related documents to be identified. Collectively, these features provide the ability to search for relevant information in a more timely and efficient manner than was previously possible. The enhanced features of the portal provide the potential to revolutionize education practice that, owing to time limitations, sometimes does not take account of research evidence at all."

Ananiadou, S., Thompson, P., Thomas, J., Mu, T., Oliver, S., Rickinson, M., Sasaki, Y., Weissenbacher, D. and McNaught, J. (2010) **Supporting the education evidence portal via text mining.** Phil. Trans. R. Soc. A, 368(1925): 3829-3844." *British Library submission*

These examples demonstrate the potential of text mining to increase the speed of processes and reduce transaction costs across a range of applications. The main advantage is to increase the potential to make discoveries by accelerating the rate at which analysis of the high volume of information deposited in online repositories can be undertaken. There is an opportunity to enhance UK competitiveness by permitting such activities, and the clarification of the law in relation to data and text mining would provide an incentive for research intensive companies, particularly in biosciences and pharmaceutical industries, to invest in the UK. The pharmaceutical sector is a strong proponent of such technologies, as evidenced by AstraZeneca's submission to the Review, which solely relates to this issue.

"This is where text mining has the potential to deliver great benefits - to industry, to patients to publishers and Great Britain plc. Analyzing the results of a therapeutic search and then text mining the large quantities of information it reveals, makes it possible to discover potential avenues for further research. By enabling researchers to identify recurring themes that are not immediately obvious – for example where a specific gene is mentioned briefly in several papers and may not stand out on its own, but collectively it becomes very significant; this can give British companies the clue that leads to an important medical advance. An excellent review of text mining and examples of successes can be found in Nature Reviews Drug Discovery.

Text mining is usually carried out by selecting a large set of references, by a very general search, and then indexing the papers retrieved to identify recurring themes. Once an interesting avenue of research has been identified the papers that mention this research are then obtained from the publisher (which also benefits the publisher) or they are licensed in advance (as is the case 98% of the time for AZ and GSK). As the pharmaceutical company already pay for access to these articles, this process of text mining does not deprive the publisher of any revenue, on the contrary it encourages greater usage of journal content and therefore helps pharmaceutical companies justify large investments in published information.

An argument publishers might make is that the downloading of large quantities of their papers, needed for indexing, might be used to misappropriate their intellectual property without payment or may be used for resale. This is not the case – it is merely a way to find and identify content, which has already been or would be purchased. Furthermore the pharmaceutical industry understands the importance of intellectual property, our business is built on IP, and we respect IP belonging to others." *AstraZeneca submission*

Opinion within the university and research sector is aligned in favour of text mining; many submissions to the *Call for Evidence* have referenced their support for one another (for example, JISC and the British Library), and there seems to be a consensus in favour of permitting such activities among wider stakeholder groups where they can be conducted without eroding intellectual property rights.

"Many modern research and archiving techniques using digital technology are not permitted under the UK's existing copyright regime. For example, data and text mining - methods central to making advances in modern medicine - are not possible with a large proportion of the digital research material available in this country. Modernising our copyright laws for teaching and research usage is a crucial element in making our universities, archival collections and libraries world class research centres and able to foster the next generation of innovators." *British Film Institute submission*

"Given the huge volumes of data and text that are now available in digital form, and that computers are able to copy and interrogate vast swathes of content we can no longer continue to rely on limitations and exceptions that essentially relate to manual "photocopying". Text-mining, data mining or media mining is the extracting of "chunks" of data using computer programmes to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modelling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow facts or hypotheses to be discovered or analysed. The technique can be used for all disciplines but perhaps offers the greatest potential in the field of medical research where it can radically increase the speed of innovation.

Algorithms are programmed to look for relationships between certain facts – for example the relationship in 3000 articles from numerous different publishers and authors between a certain enzyme and a particular cancer. In this example, once the data that relates to these two facts in the articles is extracted, a derived dataset is then interrogated further and a link, fact or hypotheses can then be evaluated. Clearly this process involves using copyright works as well as database rights, and is not routinely permitted in contracts offered to universities. Having to negotiate text and data mining clauses with numerous publishers and coming to compatible terms and conditions is likely to be an extremely long and complex process if not impossible.

Given that this new process offers enormous potential to speed up scientific discovery, and that facts (which are not copyrightable in themselves) are extracted, the argument to make text and data mining a new exception would appear to be an extremely strong one. In researching this response it has become clear that important, potentially life-saving innovation is being stymied. As stated to the Library by Dr Cameron Neylon from the Science and Technology Facilities Council "People won't/can't talk about the details because they are unsure of the legality of what they've done. In turn this means the [data and text mining] tools aren't developed because people are unsure whether they will be allowed to use them." *British Library submission*

"We believe that the emergent internet business models, the creation of open education resources, co-production of new digital assets, extensive open source software, data and text mining, mobile devices, APIs, linked data, shared services through cloud computing and other forms of "mash-up" collaborations, which extend across national geo—political boundaries provide UK innovation, education and research with unprecedented opportunities to compete internationally. And yet, these economic, social and technological innovations are endangered by the limitations of the UK intellectual property regime, which to a large extent inhibits rather than incentivise innovation and growth." *JISC submission*

“Text mining has moreover evolved into a set of key technologies that help researchers who are struggling with the problems of information overload and information overlook, due to the amount of existing research literature and the high rate of production of new literature. It goes far beyond its cousin, information retrieval, with which it is not to be confused, by providing means to rapidly drill down to individual facts, rather than, like information retrieval, providing many documents to wade through.

As text mining necessarily is applied to text, and as typically it is ideally targeted, in the sciences, at published, peer-reviewed text, and moreover at very large quantities of such text, deriving new knowledge in the process, there are severe IP issues that arise and that represent a block to innovation and progress.

...even in licences that are otherwise sympathetic to text mining, there is often a licence clause that requires individual author attribution. For example, the licence of the open access publisher BioMed Central allows text mining and the production of derivative works, but requires such attribution (this licence is identical to the Creative Commons Attribution License). In a scenario where large-scale computations are being carried out over information extracted from large numbers of documents, with massive results being fed into association mining, it becomes impossible to make such attributions in respect of some particular association, i.e. a derived piece of knowledge expressed nowhere in any text. It becomes equally impossible to contact individual authors to seek a waiver of such attribution, as allowed by the licence, given the author numbers involved. Such attribution is naturally feasible in terms of e.g. direct re-use of a single identifiable work, however this requirement effectively blocks text mining for associations over an entire archive, hence represents a barrier to knowledge discovery and innovation.

...much in fact hinges on what publishers’ licences permit. Close examination of licences reveals inconsistencies and obstructions in relation to text mining. It is likely that these are due to a long tradition of concern with physical form and content, and with derivative works that are close transforms of the original (e.g. summaries, translations). Most licences are formulated in such a way that they exclude the possibility of text mining.” *National Centre for Text Mining submission*

“Many companies active in the United Kingdom’s growing digital economy have products which could, if permitted to do so, gather, analyse and transform data in order to perform statistical analyses, prediction, data modelling, data mining, and the like. However, in the absence of well thought-out changes to the legal framework to accommodate the legitimate needs of technology companies in the United Kingdom’s economy, the United Kingdom risks being left behind in this next wave of development.” *IBM submission*

“Researchers are continually frustrated with the limited ways in which they can access and use data and information. This is certainly the case with a process known as text and data mining. Using a combination of machine learning, statistical analysis, modelling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow facts or hypotheses to be discovered and analysed. Data is selected and normalised (copied and format shifted) to allow computer programmes to analyse the data using text mining tools. Algorithms are programmed to look for relationships between certain facts across a wide range of data and information, for example 3000 journal articles from numerous different publishers and authors. Once the data that relates to these two facts in the articles is extracted, a derived dataset is then interrogated further and a link, fact or hypothesis derived.

The process of text and data mining is not routinely permitted in contracts with university researchers and may be seen to infringe copyright and database rights. As such, the process of negotiating appropriate clauses in contracts with numerous publishers to account for text and data mining would be more arduous and complex than introducing a copyright exception to allow for this activity, particularly given that text and data mining has enormous potential to speed up scientific discovery with the extraction of facts.” *The Libraries and Archives Copyright Alliance submission*

“Text and data mining, a branch of computer science that allows knowledge to be extracted from unstructured data, was originally used by businesses for purposes such as marketing and scientific research. This technology is now becoming more widely used in order to gather information from large corpora of data. Unlike web searching, text mining not only involves the retrieval of data but also its extraction and analysis in order to identify associations between different data.

Used in this way, text and data mining has great potential as a tool for retrieving and analysing data across numerous collections and institutions and also spanning a range of disciplines, cultures and languages. Despite the potential value of text and data mining software to research and innovation, there is uncertainty regarding the legality of these technologies. Contracts and licences can also restrict text mining, resulting in vast amount of knowledge being inaccessible to researchers. The introduction of an exception permitting the use of text mining technologies would clarify the situation with regard to the use of these valuable research tools.” *National Library for Wales submission*

“It must be recognised that with developments in technology, the copyright permissions regime should be capable of process improvement to serve the wider stakeholder community. That doesn’t mean that the need for permissions should be removed - but that the processes to secure permissions for reuse should be streamlined and automated. We refer to the submissions by the EPC and the UK PA on this issue.

This is already well underway. Elsevier’s 1,800 journal subscriber licences incorporate a wide range of reuse permissions as standard. These include for example, all uses required for scholarly sharing, including transmitting excerpts of articles by email or in print, to colleagues for purposes of scholarly research, whether or not those individuals are affiliated with an institute licences to access that content. It also permits members of the general public to use terminals physically located at that institution to access, search, browse, view and print journal articles at no charge through flexible licensing arrangements. We also facilitate text mining of our content to help customers deepen their insight and understanding.” *Reed Elsevier submission*

[1] Leach SM, Tipney H, Feng W, Baumgartner WA Jr., Kasliwal P, et al. 2009, *Biomedical Discovery Acceleration, with Applications to Craniofacial Development*, PloS Computational Biology 5(3): e1000215. doi:10.1371/journal.pcbi.1000215

[2] Hildebrandt M, 2010, *The Meaning and the Mining of Legal Texts*, Presentation at The Computational Turn in the Humanities. Swansea University. A further developed version will be published in: Berry, D. M. (Ed.) (forthcoming, 2011) Understanding Digital Humanities: The Computational Turn and New Technology. London: Palgrave Macmillan.

[3] Fan W, Wallace L, Rich S and Zhang Z, 2006, *Tapping the power of text mining,* CACM 4(9): 76-82. see also: http://www.computerworld.com/s/article/85113/ClearForest_Scaling_Dow_s_Paper_Mountain