Data Mining White Paper: Analysis of UK/EU law on data mining in higher education institutions

Andres Guadamuz[*] and Diane Cabell[**]

Table of Contents

---

[*] Andres Guadamuz is a consultant at Innova Technology, a software firm in Costa Rica. He is also Associate Director of the SCRIPT Centre IP and Technologies at the University of Edinburgh, where he has also served as Lecturer in Electronic Commerce Law. He has worked as an international consultant for the World Intellectual Property Organization and is currently the representative to the same body for Creative Commons. Andres has published extensively in the area of the intersection of law and technology, and has just published a book entitled "Networks, Complexity and Internet Regulation: Scale-Free Law".

[**] Diane Cabell is a Visiting Academic at the Oxford University's eResearch Center as well as Corporate Counsel for Creative Commons and Executive Director of iCommons Ltd. She served as the Associate Director of the Berkman Center for Internet & Society at Harvard where she founded the Clinical Program in Cyberlaw and remains a Fellow Emeritus. She has also served as Co---chair of the Boston Bar Association's Computer & Internet Law Committee, Visiting Scholar at the Institutt for informatikk Universitetet in Oslo, and Assistant Counsel and Faculty Resident at the Massachusetts Institute of Technology.

Acknowledgments

This paper is distributed under the Creative Commons Attribution 3.0 License (CC BY). It has been prepared for Wikipedia founder Jimmy Wales in advising the Universities and Science Minister David Willetts on the terms of access to the proposed Gateway to Research project. See http://bit.ly/Ry0FWU.

The authors would like to thank Kusuma Trust UK for its generous support of the iCommons Open Collaboration Research Project without which this paper would not be possible. The authors would also like to thank Dr. Abbe Brown, Senior Lecturer at the University of Aberdeen, Dr. Dinusha Mendis, Senior Lecturer at Bournemouth University, Dr. Prodromos Tsiavos, adviser on legal issues of open data in the Greek Prime Minister's e-Government Task Force and the Special Secretary for Digital Planning, and Diane Peters, Creative Commons General Counsel for their helpful input.

1. Introduction

Data or text mining (hereafter called "content mining") is a process that uses software that looks for interesting or important patterns in data that might otherwise not be observed. An example might be combining a database of journal articles about ground water pollution with one of hospital admissions to detect a pollution-related pattern of disease breakout.

It is also a useful tool in commerce. A credit card company might detect a correlation between purchases of tickets from particular airline with purchases of certain types of automobiles and develop a marketing program uniting appropriate vendors. One McKinsey report states that the utilization of 'big data' in the sphere of public data alone could create €250 billion annual value to Europe's economy.[1]

Content mining is increasingly accomplished by machine. Databases, particularly those produced by scientific research, are far too large to be scanned by human eyeball. However, the right to mine data is not assured by the law in most jurisdictions and even where it is, the terms of access to the majority of research publication databases deny permission to do so. One recent study indicated that obtaining permission to mine the thousands of articles appearing on a single subject from the myriad of different publishers would require 62% of a researcher's time. Many content owners, including research institutions, have yet to develop any policy on content mining.[2].

This report will identify the main legal barriers to data mining and data reuse and make policy suggestions to guide governments, funding agencies, and research institutions. As the title suggests, the emphasis of the study is about legal issues that are specific to higher education institutions (HEIs).

The first challenge for this report is to attempt to delimit the subject matter, as various types of content that are subject to automated analysis.[3] HEIs can hold and share content of various formats, here are just a few examples:

---

[1] McKinsey Global Institute, *Big Data: The next frontier for innovation, competition and productivity*, (2011).

[2] McDonald, *Value and benefits of text mining*, March 2012 at http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

[3] Research Information Network, *Stewardship of Digital Research Data - Principles and Guidelines*. London: RIN (2008), http://www.rin.ac.uk/data-principles.

- **Text:** published articles, book chapters, preparatory notes, working papers, reports, teaching materials, conference papers, presentations, theses.
- **Datasets:** statistical data, geolocation data, survey results, maps, figures, time series, genetic information, health records, computer logs.
- **Multimedia:** pictures, sound recordings, interviews, presentations, video.

Each of the above may have separate legal regimes applying to them. In the interest of convenience and simplicity, whenever the report talks about database contents, there will be no distinction as to whether we are dealing with text, data or multimedia, unless clearly specified in the text.

2. Content mining

It is an undeniable fact that databases are growing in number and size.[4] This increase in data has prompted a change in the way in which we look at large datasets, as it becomes impossible for humans alone to sift through new knowledge. As a response to this challenge, computational technologies and techniques are increasingly used to retrieve and analyse data held in something called "knowledge discovery in databases" (KDD). Data mining is a subset of this branch of data analysis. While it may not be perfect, the mining analogy serves to explain roughly what content mining entails. Artificial intelligence agents sift through large amounts of data, eventually finding valuable information which was undiscovered before. Moreover, in large mining operations one sifts through large quantities of low-grade material in order to find something valuable.

As explained by Fayyad et al:

> *KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.*

For the purposes of the present report, content mining is to be described as the extraction of data from large datasets to uncover previously unknown and potentially useful information.[5] While the field is relatively new, increased computing capabilities make the analysis of large datasets not only possible, but useful. The applications for content mining range from the mundane to the transcendental. For example, studies have used text mining techniques to explore social sentiment[6] and public opinion[7] through the analysis of social media. Other studies have been looking at the use of social media to survey health and disease occurrences, for example, by looking for the prevalence

[4] Fayyad U, Piatetsky-Shapiro G, and Smyth P, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* 37 (1996).

[5] Frawley WJ, Piatetsky-Shapiro G, and Matheus CJ, "Knowledge Discovery in Databases: An Overview", 13:3 *AI Magazine* 57 (1992).

[6] Pang B and Lee L, "Opinion Mining and Sentiment Analysis", 2:1 *Foundations and Trends in Information Retrieval* 1 (2008).

[7] O'Connor B et al, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010).

of mentions of influenza online.[8] More serious applications include the use of content mining in biology and medicine.[9]

The methods for extracting and analysing the data may be relevant for the legal questions that are the subject of this report. There are various types of content mining, for example, some look at anomalous records, or look for correlations and/or dependencies in the data. These techniques use different software and algorithms, so it is difficult to generalise for legal purposes. However, the statistical analysis usually associated with content mining requires access to the data, and the possibility of creating some form of remote copy for analysis purposes (although actual copies are not always necessary). Similarly, the analysis of the data tends to be aggregated and reused to produce tables, diagrams and histograms of the combined sets.[10]

It is difficult to generalise on what exactly is the method for content mining, as there are different algorithmic and model structures depending on the subject, the type of database, and the type of analysis being performed.[11] For the purpose of this study, it will be assumed that most content mining roughly follows these steps (Figure 1):

1. Individual content is created.
2. Content is placed into data set, repository or collection.
3. Miner gains access to the data.
4. Mining tools applied to the data set.
5. Analysis of the processed data.
6. New knowledge.[12]

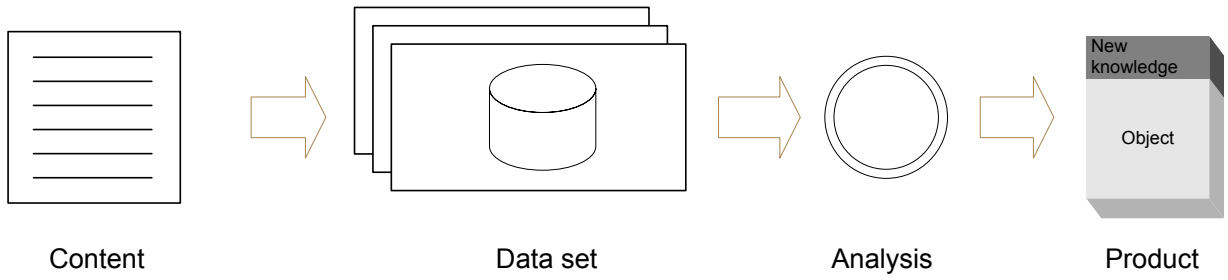

Figure 1. A typical content mining operation.

[8] Corley C et al, "Text and Structural Data Mining of Influenza Mentions in Web and Social Media", 7:2 *International Journal of Environmental Research and Public Health* 596 (2010).

[9] See for example Krallinger M, Valencia A and Hirschman L, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology", 9:2 *Genome Biology* S8 (2008); and Ananiadou1 S, Kell DB, and Tsujii J, "Text Mining and its Potential Applications in Systems Biology" 24:12 *Trends in Biotechnology* 571 (2006).

[10] Han J and Kamber M, *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann Publishers (2000), p.16.

[11] Ibid, p.23.

[12] These steps are a simplified version of the processes described here: Korn N, Oppenheim C and Duncan C, IPR and Licensing issues in Derived Data, JISC report (2007), http://bit.ly/TEmtMX.

The key points from a legal perspective are stages 3 and 4. Researchers must be able to have access to the data in a format that is susceptible of analysis, for which it must be assumed that the content is either freely available, or the researcher has some form of licensing agreement. Then, there is the vital question of what operation is performed on the data. Is there copying of the entire content of the database? If not, what sort of operation is performed? Is there some form of retrieval of key data? Is the operation simply looking at patterns? What is the format of the new knowledge?

The answer to these questions may prove vital in answering the legality of content mining operations. In the interest of a general legal analysis, it will be assumed that there is actual copying of substantial sections of contents during the mining operation, although it is understood that this may not always be the case. It will also be assumed that the analysis operation means that the work has been extracted in the meaning of the database right, although this may also be open to interpretation.

3. The law

Databases are protected in the UK through a variety of norms, and each may have a bearing on the legality of content mining. Here is a list of applicable legislation.

3.1 Copyright

The data contained in databases can be protected under copyright law as a literary work. Section 3A of the Copyright, Designs and Patents Act 1988 (CDPA), defines a database as a collection of independent works which "are arranged in a systematic or methodical way", and "are individually accessible by electronic or other means". However, the threshold of originality in a database is quite high. Section 3A states that:

> *For the purposes of this Part a literary work consisting of a database is original if, and only if, by reason of the selection or arrangement of the contents of the database constitutes the author's own intellectual creation.*

This means that in UK copyright law the author's own skill and labour is required in the selection and arrangement of the contents of a database, a mere gathering of data without meeting this requirement is not worthy of protection because it does not meet the originality test. This means that mere compilations of works do not meet the standard of copyright protection.[13] It is important to stress as well that what is protected is the database as a whole, as individual elements may or may not be protected on their own.[14]

UK and European case law serve to illustrate the higher originality threshold in databases. In the English case of *Navitaire v Easyjet,*[15] Pumfrey J had to consider whether a computer-based database is a computer program or a database for copyright purposes, and interestingly found that the addition and removal of datasets, schemas and other structural changes to the arrangement of a database were to be considered computer programs instead of databases in their own right. The meaning of this ruling for databases is that there would be a protection of the source code in the shape of a literal work, and not of the functional elements as such, which are an important and

---

[13] MacQueen HL, Laurie GT and Waelde C, *Contemporary Intellectual Property: Law and Policy*, Oxford: Oxford University Press (2008), p. 66.

[14] OutLaw, *Database Rights: The Basics* (2008), http://www.out-law.com/page-5698.

[15] *Navitaire Inc v Easyjet Airline Co. & Anor* [2004] EWHC 1725 (Ch).

integral part of a database. The case spells out this dichotomy when Pomfrey J states clearly that "*Copyright protection for computer software is a given, but I do not feel that the courts should be astute to extend that protection into a region where only the functional effects of a program are in issue."*[16]

Another European case, *Football DataCo,*[17] involved the fixture lists of football matches in the English and Scottish leagues, which are produced by a company called Football DataCo. Web aggregator Yahoo! copied them without paying licence fees, so Football DataCo sued them alleging that by doing so Yahoo! had infringed both copyright and its database rights. The Court of Appeal of England and Wales referred[18] the case to the European Court of Justice (ECJ), which decided that copyright can only be afforded to a database if its structure is the maker's own intellectual creation. This continued to set a bar high of not only originality, but of the skill and labour[19] required to have protection under copyright for a database. The ECJ opined that "the significant labour and skill required for setting up that database cannot as such justify such a protection if they do not express any originality in the selection or arrangement of the data which that database contains."[20]

Assuming copyright in the database exists, regardless of the high protection threshold, then the author would have the exclusive right to authorise use and reuse of the data, and any such unauthorised use would be a copyright infringement. Acts which infringe copyright might still fall under an exception or limitation, which in the UK take the shape of fair dealing. Only those acts listed under the CDPA can be considered exceptions. Section 50D does contain a fair dealing provision with regard to databases. It reads:

> *(1) It is not an infringement of copyright in a database for a person who has a right to use the database or any part of the database, (whether under a licence to do any of the acts restricted by the copyright in the database or otherwise) to do, in the exercise of that right, anything which is necessary for the purposes of access to and use of the contents of the database or of that part of the database.*

Unfortunately, this is a very narrow exception is unlikely to cover the type of reuse of the information that is typical of content mining. Fair dealing in databases covers only those acts that are necessary to use the contents of the database, and in the strictest sense, one could argue that content mining is not a "necessary" use of the data, as the above exception seems to give permission on the basis of operational uses. Therefore, only functional uses could be considered non-infringing.

Similarly, content mining does not seem to fall under any other research-related fair dealing, as these also tend to be very narrow. For example, s29 CDPA states that:

> *(1) Fair dealing with a literary, dramatic, musical or artistic work for the purposes of research for a non-commercial purpose does not infringe any copyright in the work provided that it is accompanied by a sufficient acknowledgement.*
>
> *(1A) Fair dealing with a database for the purposes of research or private study does not infringe any copyright in the database provided that the source is indicated.[…]*

---

[16] At para 94.

[17] *Football Dataco Ltd and Others v Yahoo! UK Ltd and Others* C-604/10.

[18] [2010] EWCA Civ 1380.

[19] Also known as sweat of the brow in other jurisdictions.

[20] C-604/10 at para 46.

> *(1C) Fair dealing with a literary, dramatic, musical or artistic work for the purposes of private study does not infringe any copyright in the work.*

Any content mining operation that copies text would fall under this exception if it is for non-commercial purposes only, or if it is performed with the purpose of "private study". The definition clearly implies that content mining of medical texts by a pharmaceutical company looking for new drug treatment would clearly be an infringement, while content mining performed by an academic would find itself in more of a grey area. The problem with the research and private study exception is that, as Cornish points out, the courts have not been asked to ascertain how much can be taken, and what constitutes non-commercial use exactly.[21] The provisions can be interpreted in light of the InfoSoc Directive,[22] which in Art 5(b) contains a more comprehensive definition of what is to be considered as fair dealing for research; it reads:

> *…in respect of reproductions on any medium made by a natural person for private use and for ends that are neither directly nor indirectly commercial, on condition that the rightholders receive fair compensation which takes account of the application or non-application of technological measures referred to in Article 6 to the work or subject-matter concerned*.

It could be argued that academic research might fall under indirectly commercial use under some circumstances. Similarly, the request that rights holders should receive fair compensation denotes the restrictive interpretation given to the exception. Furthermore, content mining does not appear to fall under the exception for observing, studying and testing of computer programs (s 50BA).

The absence of a specific exception for content mining seems to indicate that if a database has copyright, most types of unauthorised content mining could be copyright infringement.

3.2 Database right

In addition to copyright protection for databases, the UK has implemented a sui generis right arising from the European Database Directive,[23] enacted in the UK through the Copyright and Rights in Databases Regulations 1997 (CRDR). It is important to point out that the database right exists regardless of the existence of copyright protection in the database, as the exclusive rights given to the database owner are separate to those arising from copyright.[24]

The database right is an exclusive right given to the maker of a database,[25] which is defined as a collection of independent works, data or other materials which are arranged in a systematic or methodical way, and are individually accessible by electronic or other means.[26] The right exists if "there has been a substantial investment in obtaining, verifying or presenting the contents of the

---

[21] Cornish WR and Llewelyn D, *Intellectual Property : Patents, Copyright, Trade Marks & Allied Rights*, 7th ed ed, London: Sweet & Maxwell (2010), p.509.

[22] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

[23] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[24] s13 CRDR.

[25] s14 CRDR.

[26] s6 CRDR.

database".[27] The right subsists for 15 years from the completion of the same.[28] The right is infringed if a person without authorisation "extracts or re-utilises all or a substantial part of the contents of the database".[29] The right is also infringed after continuous extraction or re-utilisation of non-substantial parts of the database.[30] For the purpose of the CRDR, re-utilisation is understood as making the contents of the database available to the public by any means.[31]

The database right comes with a fair dealing provision stating that there is no infringement if a substantial part has been extracted[32] or re-utilised if:

> *(a) that part is extracted from the database by a person who is apart from this paragraph a lawful user of the database,*
>
> *(b)it is extracted for the purpose of illustration for teaching or research and not for any commercial purpose, and*
>
> *(c) the source is indicated.*[33]

It is clear that the database right, if it exists in a database, precludes many forms of unauthorised content mining operations. The fair dealing provision cited above applies only if the person performing the content mining is already a lawful user of the database, the operation is done with attribution, and for research-related non-commercial purposes. We encounter here the same problem about the lack of definition of what constitutes non-commercial use. It may be advisable to interpret this provision also in light of the InfoSoc Directive, as was done in the previous section with regards to copyright. This would mean that any direct or indirect commercial use might be infringing. For example, an academic who is funded by a pharmaceutical company for his research at the university might fall outside of what is permitted under fair dealing.

However, the ECJ delivered a set of decisions that watered down the database right by raising the bar of what databases can be said to meet the standard of protection. In 2004, the ECJ delivered a number of decisions clarifying the database right, of which one was a referral from an English court. In *British Horseracing Board v William Hill*[34], the ECJ was asked to determine whether the collection of horse racing information obtained through a third party by the defendants was a database subject to the sui generis right. The betting agency William Hill obtained horse racing data by a licensing agreement with a third party, not with the British Horseracing Board, which created the data. While most of the case rested on the issue of whether there had been substantial extraction of data from the original, an important part of the decision was in regard to whether the database maker had incurred enough investment to warrant protection. Here the court decided that:

> *The expression 'investment in … the … verification … of the contents' of a database in Article 7(1) of Directive 96/9 must be understood to refer to the resources used, with a*

---

[27] s13 CRDR.

[28] s17 CRDR.

[29] s16 CRDR.

[30] Ibid.

[31] Bently L and Sherman B, *Intellectual Property Law*, 3rd ed, Oxford: Oxford University Press (2008), p.303.

[32] Or the continuous extraction of a non-substantial part as per s16 CRDR.

[33] s20 CRDR.

[34] British Horseracing Board Ltd v William Hill Organization Ltd (BHB decision) C-203/02.

> *view to ensuring the reliability of the information contained in that database, to monitor the accuracy of the materials collected when the database was created and during its operation. The resources used for verification during the stage of creation of materials which are subsequently collected in a database do not fall within that definition.*[35]

The above paragraph seems harsh, as in it the ECJ seems to seriously erode database protection by setting a high standard of protectable investment. The paragraph is particularly severe when it comes to the investment in verifying information that goes into a database. Here the ECJ further comments:

> *…although the search for data and the verification of their accuracy at the time a database is created do not require the maker of that database to use particular resources because the data are those he created and are available to him, the fact remains that the collection of those data, their systematic or methodical arrangement in the database, the organization of their individual accessibility and the verification of their accuracy through the operation of the database may require substantial investment in quantitative and/or qualitative terms within the meaning of Article7(1) of the Directive.*[36]

This means that the ECJ has not done away with verification altogether, it simply establishes high level of investment in all of those steps is required. As many commentators have noted, this significantly reduces the potential scope of the database right, as only those databases that meet the higher standard of investment are protected.[37]

The result of the ECJ ruling is difficult to ascertain, but it is increasingly likely that the database right has not met the initial expectations for which it was created. The European Commission conducted a review of the impact of the new right, and found that it had no effect whatsoever in fostering the creation of a new sector in the European economy. In 1996, the United States (which provides no sui generis database protection) had the largest share of the global database market, with 56%, while European share was 22%. While this share increased between 1996 and 2001, it had dropped again to 24% by 2004, while the U.S. share went back to its previous levels.[38] This is strong indication that the sui generis right did not have any noticeable effect in strengthening the European database market. In an indicting comment on policy based on lobbying and guesswork, the Commission's report said:

> *Nevertheless, as the figures discussed below demonstrate, there has been a considerable growth in database production in the US, whereas, in the EU, the introduction of "sui generis" protection appears to have had the opposite effect. With respect to "non-original" databases, the assumption that more and more layers of IP protection means more innovation and growth appears not to hold up.*[39]

---

[35] Ibid at para 31.

[36] Ibid at para 36.

[37] Davison MJ , Hugenholtz PB, "Football fixtures, horse races and spin-offs: the ECJ domesticates the database right", 3 *European Intellectual Property Review* (2005).

[38] European Commission, *First Evaluation of Directive 96/9/EC on the Legal Protection of Databases*, DG Internal Market Working Paper, http://is.gd/DsY3XV.

[39] Ibid, p.24.

Despite this, there are no plans to scrap the sui-generis right.

3.3 Public Sector Information

Academic institutions are major producers of research data. As most higher educational institutions in the UK receive public funds in one way or another, it is necessary to cover the relevant norms that rule the use and reuse of public sector data. The regime in place was enacted by the 2003 Public Sector Information (PSI) Directive,[40] which has been implemented in the UK in the Re-Use of Public Sector Information Regulations 2005.[41] The purpose of the PSI system is to encourage the reuse of public sector information. Although neither the Directive nor the Regulations require public sector organisations to make documents available to the public, if they do so it should be in line with the notions of transparency, fairness and consistency.

The PSI Regulations establish an exhaustive list of institutions that are considered public sector bodies and therefore covered by the legislation. Educational institutions are specifically exempted from the Regulations in s 5(3)(b), which reads:

> *These Regulations do not apply to documents held by— […]*
>
> *(b)educational and research establishments, such as schools, universities, archives, libraries, and research facilities including organisations established for the transfer of research results;*

This exclusion is somewhat unfortunate because an important part of the UK's strategy has been the creation of a unified licensing scheme for public sector information, more of which will be covered below.

3.4 Other relevant legislation

Depending on the type of database, some other legislation could possibly be applicable to content mining.

Data protection could be of concern when mining databases that might contain personal data, but more importantly, sensitive personal data. This covers, amongst other, data which contains a subject's racial or ethnic origin, political opinions, religious beliefs, health records, and sexual life.[42] Those who process such data are considered data controllers and should follow the data protection principles[43] mandated in the Data Protection Act 1998, but also should notify the Information Commissioner that they are indeed processing personal data.

Another legislation that may apply to content mining is the INSPIRE Directive,[44] which sets an obligation for public authorities which hold spatial and location-based data to make it available in consistent formats through networked services. These services must make possible "to search for

---

[40] Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information.

[41] The Re-use of Public Sector Information Regulations 2005, SI No. 1515.

[42] s 2 Data Protection Act 1998.

[43] For a quick guide to the DP principles, see:
http://www.ico.gov.uk/for_organisations/data_protection/the_guide.aspx.

[44] Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE).

spatial data sets and spatial data services on the basis of the content of the corresponding metadata and to display the content of the metadata".[45] The relevance to content mining is that, unlike the PSI Directive, the INSPIRE framework does not exclude specifically educational and research institutions. On the contrary, the definition of what is considered a public authority subject to the regulations is rather broad, and includes bodies that perform some form of public administration that runs a spatial data service or holds spatial data.[46]

4. Open Access Policies

The UK is fast becoming one of the most forward-looking countries with regard to opening access to research, in part thanks to a shift in policy from funding bodies in favour of wider access to research, but also due to growing government pressure in that respect. The rise of open access[47] in higher education institutions is of great importance for content mining as it can free up databases and other resources to analytical exercises. This is particularly relevant because, as we have seen before, these works may be restricted either by copyright or by the database right.

Significant pressure to make research more openly available has come from investigators themselves, with prominent academic voices coming out in favour of open access.[48] One such example is the Manchester Manifesto,[49] a document drafted by UK and European scientists trying to answer the question "who owns science?" They conclude that:

> *Scientific information, freely and openly communicated, adds to the body of knowledge and understanding upon which the progress of humanity depends. Information must remain available to science and this depends on open communication and dissemination of information, including that used in innovation.*

Another valuable pillar in the success of open access has been the fact that funding bodies are increasingly requiring that any research they support financially must be released at some point to the public, be it via institutional repositories, self-publishing, or through other similar means. The Wellcome Trust has enacted an Open Access Policy which makes it clear that, while it expects funded research to be published in peer-reviewed journals, it also requires that such works should eventually be made available to the public for free through PubMedCentral UK[50] within six months of publication.[51] Similarly, Research Councils UK, the partnership of the seven higher education funding research councils, has also established an updated open access policy[52] which states that all

---

[45] s 7(2)(a)(i) INSPIRE Regulations 2009.

[46] s 3 INSPIRE Regulations 2009.

[47] It is assumed that the reader is already familiar with open access. If that is not the case, the Berlin Declaration on Open Access defines it as "a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community. […] Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material." See: http://is.gd/HTZLr6.

[48] Mathemetician Tim Gower boycott against Elsevier; Mark Walport or other signatories to Bethesda Statement on Open Access Publishing

[49] Addison T et al, *The Manchester Manifesto,* Institute for Science, Ethics and Innovation (2009).

[50] Soon to be Europe PubMed Central.

[51] See the policy here: http://is.gd/rHhQM9.

[52] RCUK's 2012 policy version can be found here: http://is.gd/xbjUDv.

publicly-funded research must be published in an open access journal that allows "immediate and unrestricted access to the publisher's final version of the paper". If the journal does not offer such an option, then the work must be published in a journal that allows the work to be placed in other repositories "without restrictions on non-commercial re-use and within a defined period". Such clear and unequivocal statements in support of open access are transforming scientific publishing, and allow more works to be accessible for mining.

The UK government itself has also been directly responsible for encouraging wider adoption of open access. One of the main drivers of this push has been the Joint Information Systems Committee (JISC), which is an independent quasi-autonomous non-governmental organisation (QUANGO) supported by the main national higher education funding councils and by the Department for Employment and Learning. Its main role has been to support and finance internal and external projects related to all aspects of information management in education, including projects on digital repositories, archives, content mining, preservation, metadata, standards, and interoperability. In exercising this function, JISC has produced a considerable number of reports in favour of open access,[53] but it has also created a substantial infrastructure that provides tools necessary for open access.

An important part of the work of JISC when it comes to open access has been to promote and encourage researchers in HEIs to upload content to institutional repositories. Needless to say, this is a vital part of any open access strategy. Besides having published guides on how to promote the adoption and use of repositories,[54] JISC has funded projects that try to find ways in which to encourage open access.[55]

In November 2010, the government commissioned an independent review on how intellectual property supports growth and innovation. The Hargreaves Review of Intellectual Property[56] produced a series of interesting and balanced recommendations. The study specifically mentions text mining as a subject that requires a new exception in copyright. The Review states:

> *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. The Government should introduce a UK exception in the interim under the non-commercial research heading to allow use of analytics for non-commercial use, as in the malaria example above, as well as promoting at EU level an exception to support text mining and data analytics for commercial use.*[57]

The current UK government administration has indicated its support for Hargreaves' recommendations in the belief that it will not only stimulate scientific research but will also enable greater commercialization of UK know-how. This potential was also recognized in the latest and most comprehensive review on open access, the Report of the Working Group on Expanding Access

---

[53] For some reports, see: http://is.gd/NKvBIb.

[54] See for example: http://bit.ly/NPPzI2.

[55] See for example: Proudfoot RE et al, J*ISC Final Report: IncReASe (Increasing Repository Content through Automation and Services)*, White Rose Consortium (2009).

[56] Intellectual Property Office, *Digital Opportunity: A Review of Intellectual Property and Growth*, (2011), http://www.ipo.gov.uk/ipreview.htm.

[57] Ibid, para 5.26.

to Published Research Findings (Finch Report).[58] The group was established by the Minister for Universities and Science in the context of the Research Innovation Network, and was tasked with advising the government on it policies with regards to scientific research. Although the Report does not study content mining in depth or suggest any other solutions beyond those of the Hargreaves Review, the report comments:

> *Related to such moves has been a growth of interest in exploiting the potential of text-mining tools to analyse and process the information contained in collections or corpora of journal articles and other documents in order to extract relevant information, to manipulate it, and to generate new information. The use of such techniques is not yet widespread, not least because arrangements for making publications available for text mining can be complex, and because the entry costs are high for those who lack the necessary technical skills. But text mining offers considerable potential to increase the efficiency, effectiveness and quality of research, to unlock hidden information, and to develop new knowledge.*[59]

The Finch Report came out strongly in favour of open access as a matter of government policy, encouraging OA publishing through article processing or publishing charges (APC)[60] whereby the expense of publication in an open access journal is borne by the grantee research institution, whenever there have been public funds have been used in the research. Similarly, it advises that an effective public policy towards open access should be accompanied by an effort to "minimise restrictions on the rights of use and reuse, especially for non-commercial purposes, and on the ability to use the latest tools and services to organise and manipulate text and other content".[61] Although Finch's preference for the author-pays model (so-called "gold" open access) [as opposed to the "green" OA method which allows authors to self-publish the work in any open access repository] has prompted some criticism,[62] there can be little doubt that the above constitutes a fundamental shift in favour of future access to research, including access to reuse by content mining.

Even more encouraging is the announcement by the government that it will be implementing the Finch Report's recommendations. Furthermore, they have guaranteed that all future research funded by public money will be available without restrictions anywhere in the world.[63] Finally, open access advocates have started to campaign in earnest in favour of content mining of scholarly publications. In a recent article, molecular scientist and OA expert Peter Murray-Rust formulated the concept of "open content mining", defining it as:

> *… the unrestricted right of subscribers to extract, process and republish content manually or by machine in whatever form (text, diagrams, images, data, audio, video,*

---

[58] *Accessibility, sustainability, excellence: how to expand access to research publications*. Report of the Working Group on Expanding Access to Published Research Findings: http://is.gd/91tsKb.

[59] Ibid, para 3.19.

[60] Various terms are used to define this work

[61] Ibid, p.7.

[62] Ayris P, "Why panning for gold may be detrimental to open access research", *The Guardian* (23 July 2012), http://is.gd/uscUS3.

[63] Sample I, "Free access to British scientific research within two years", *The Guardian* (15 July 2012), http://is.gd/yOCTus.

etc.) without prior specific permissions and subject only to community norms of responsible behaviour in the electronic age.[64]

In the article he proposes three main principles governing open content mining. These are:

1. *Right of Legitimate Accessors to Mine.* There should be no objection to automated analysis of published works in the interest of research.
2. *Lightweight Processing Terms and Conditions.* Licensing and other terms and conditions should not restrict mining.
3. *Use.* Researchers should be able to publish and disseminate the result of their analysis.

These principles are a sign of the growing importance of content mining, but are also a welcome addition to the intellectual and ethical push towards more open research environment.

5. Licensing[65]

Until the open access government recommendations are fully implemented and assuming that a database is protected by copyright and/or the database right, then content mining can be performed legally only with adequate permission to do so. This is where the terms and conditions governing data use and reuse require careful analysis. . If we are thinking of higher education data, it should be held in a repository or archive of some sort. While these will be covered in more detail later, it is important to enumerate possible licensing schemes under which databases are already offered, or under which they could be released in the future.

5.1 Creative Commons

The most prevalent[66] open access licences are those offered by Creative Commons (CC) which is a non-profit organisation founded in 2001 in the US with the aim of promoting science and the arts by making it easier for authors and creators to offer a flexible range of protections and freedoms to users of their works. It counters the "all rights reserved" tradition associated with copyright by introduction a set of licences in which authors keep only "some rights reserved". These licences range from dedicating the work straight to the public domain, to more narrow licences with several restrictions.

There are several versions of the licences, from CC 1.0 to the latest version 3.0. At the time of writing, there is a drafting process in place to update the licence to version 4.0. Besides these numbered versions, the licences have been ported to comply with local legislation in over 50 jurisdictions, and are in process of localization in over 20 more countries. In the UK, there are two versions of CC licences for the two main jurisdictions, version 2.0 for England and Wales, and version 2.5 for Scotland. This means that some authors may prefer using the unported general 3.0 version.

---

[64] Murray-Rust P, "The Right to Read Is the Right to Mine", *Open Knowledge Foundation Blog* (June 1, 2012), http://bit.ly/O75Rwd.

[65] Disclaimer: The author is Legal Lead for Creative Commons Costa Rica, Liasion to the World Intellectual Proeprty Organization for Creative Commons, and also serves in the Open Data Commons Advisory Council.

[66] By the end of 2010, there were 400 million works released under a CC licence. See: http://wiki.creativecommons.org/Metrics.

All Creative Commons licences (excepting CC0, which is a public domain dedication and therefore not strictly a licence) work with copyright protection by maintaining a minimum set of standards met by all of their offered legal documents. The licences grant users the right to reproduce, distribute, publicly perform and make modifications. In exchange, all licencees have to meet several common conditions. These include:

- The user must attribute the work in any reproduction or redistribution of the work. This is known as the Attribution licence element (BY), and it is common in all licences after version 2.0.
- Fair use rights, fair dealing, or any other acquired exceptions are not affected by the licence.
- Copyright notices should not be removed from all copies of the work.
- Every copy of the work should maintain a link to the licence.
- Licensees cannot use technological protection measures to restrict access to the work.
- The licences have worldwide application, have lasts for the entire duration of copyright (unless otherwise specified), and are irrevocable.

Besides these rights and restrictions, licensors can choose to add up to three additional licence elements:

- Non-commercial (NC): The work can be copied, displayed and distributed by the public, but only if these actions are for non-commercial purposes.
- No derivative works (ND): This licence grants baseline rights, but it does not allow derivative works to be created from the original.
- Share-Alike (SA): Derivative works can be created and distributed based on the original, but only if the same type of licence is used.

ND and SA are exclusive, so this means that there are 6 possible CC licences mixing and matching those elements. These are:

- Attribution (BY)
- Attribution - Non Commercial (BY-NC)
- Attribution - Share Alike (BY-SA)
- Attribution - No Derivatives (BY-ND)
- Attribution - Non Commercial - Share Alike (BY-NC-SA)
- Attribution - Non Commercial - No Derivatives (BY-NC-ND)

The most restrictive licences are evidently BY-NC-ND and BY-NC-SA, while BY would be the one which allows more reuse possibilities, including commercial use. All CC licences are presented in three formats: the first is a short and easy to read “Commons Deed”, which explains the terms and conditions of the licence in a simple manner; the second format is the “Legal Code”, which is the full licence; the third is the “Digital Code”, which provides a machine-readable version of the licence in RDF[67] format.

It must be pointed out that the licence specifically allows users to individually negotiate terms and conditions in order to obtain specific permission from the author to perform one of the acts

[67] Resource Description Framework (RDF) is a metadata format.

restricted by the licence. For example, if the licence does not allow modification of a work, this action could only be performed with the permission of the owner.

Besides the described licences, Creative Commons also offers creators the possibility of dedicating their work to the public domain via a document called CC0 (CC Zero).[68] However, dedications to the public domain seem to have a difficult status in law, as this is not something that is contemplated in copyright treaties, so it is possible that a full release of a work to the public domain before copyright has expired may not be possible in some jurisdictions.[69] Particularly for the UK, there is a strong case to be made that works under copyright cannot be unilaterally placed into the public domain; Johnson tried finding any legal authority in both English and Scots law that copyright can be dedicated to the public domain, and found none.[70] Because of this, CC0 acts as a waiver of existing rights, where the authors express that, to the fullest possible extent of the law, they will not enforce their rights. These rights include all copyright on the work, but also list related rights, which include specific mention of data and the database right. CC0 says that the following fall under the definition of rights waived:

> *v. rights protecting the extraction, dissemination, use and reuse of data in a Work;*
>
> *vi. database rights (such as those arising under Directive 96/9/EC of the European Parliament and of the Council of 11March 1996 on the legal protection of databases, and under any national implementation thereof, including any amended or successor version of such directive);*

In case the above is not allowed because of legal prohibitions against waivers, CC0 contains a fall-back licence that grants a "a royalty-free, non transferable, non sublicensable, non exclusive, irrevocable and unconditional license to exercise" all copyright on the work, which is has the same effects as if the work was not protected by copyright.

There may be some problems with using Creative Commons to release databases in the UK. Firstly, with the exception of the aforementioned CC0, Creative Commons licences are mainly copyright-related documents and do not specifically mention the database right; thus they may not be deemed applicable in countries subject to the Directive.[71] Secondly, as stated above, the two UK licences in existence are not the latest version, 2.5 in Scotland and 2.0 for England and Wales. While there is no reason to believe that version 3.0 is invalid in the UK, some institutions may think twice before using an international port.

However, these fears may be laid to rest when CC finalises and releases version 4.0 of its licensing suite. This will have some interesting features, firstly, it will be an international licence, meaning that CC will minimize the need to create country-specific ports. Furthermore, if the published draft is anything to go by, CC 4.0 will specifically protect databases, including the database right. The draft 4.0d2[72] now grants permission to use and reuse the work subject to copyright or Copyright-

---

[68] Text here: http://creativecommons.org/publicdomain/zero/1.0/.

[69] See particularly: Dusollier S, *Scoping Study on Copyright and Related Rights and the Public Domain*, Study for WIPO Committee on Development and Intellectual Property (CDIP/7/INF/2).

[70] Johnson P, "Dedicating Copyright to the Public Domain", 71:4 *Modern Law Review* 587 (2008).

[71] CC takes the position that it covers the sui-generis database rights, although they are not specifically mentioned. The right to copy would include the right to extract; the right to adapt covers re-utilization, and so on.

[72] CC 4.0 drafts can be found here: http://wiki.creativecommons.org/4.0/Drafts.

like Rights. The licence defines Copyright-like Rights as "those rights that neighbor or are similarly related to copyright, such as performance, broadcast, phonogram and database rights, without regard to how such rights are named, labelled or categorized." This should make CC a perfectly viable option for licensing databases.

5.2 Open Data Commons

The Open Data Commons is a set of licences and dedications created by the Open Knowledge Foundation (OKF) that are specifically directed towards protecting databases. The project was started as an independent work by Jordan Hatcher and Prof. Charlotte Waelde in 2007 and funded by the software company Talis. This first effort produced the Open Database Licence (ODbL),[73] and then the project was transferred to the OKF in 2009. An advisory board was convened and one more licence and one dedication were added, the Open Data Commons Attribution License,[74] and the Open Data Commons Public Domain Dedication and License (PDDL).[75]

The project was started because the drafters noticed that Creative Commons was not covering the database right specifically which they believed left some institutions in Europe at potential risk due to market failure as they could licence only their copyright and not the database right. It was therefore felt that a database specific licence was needed.

As previously mentioned, the ODbL covers the database right, but it also licenses copyright. Interestingly, while this strongly implies that the licence is applicable only within European jurisdictions that have the sui generis right, the licence specifies that it is also a contract between the licensor and the user. The effect of this small legal trick is that it allows the licence to extend the effects of the database right to jurisdictions where it does not exist through share-alike clauses, as the protection will therefore be contractual. The licence grants the following rights:

> a. Extraction and re-utilisation of the whole or a substantial part of the contents.
>
> b. Creation of a derivative database; e.g. this includes any translation, adaptation, arrangement, modification, or any other alteration of the database or of a substantial part of the contents.
>
> c. Inclusion of the database in unmodified form as part of a collection of independent databases.
>
> d. Creation of temporary or permanent reproductions by any means and in any form, in whole or in part.
>
> e. Distribution, communication, display, lending, making available, or performance to the public by any means and in any form.

In exchange, the user must fulfil several conditions. These include the obligation to keep copyright and database notices intact, and this being a share-alike licence, the user must release any derivatives under the terms of the ODbL. The user is also forbidden from releasing derivatives imposing any form of technological protection measure. Most of the other provisions in the licence are similar to those found in CC licences.

---

[73] Full text here: http://opendatacommons.org/licenses/odbl/1.0/.

[74] Full text here: http://opendatacommons.org/licenses/by/.

[75] Full text here: http://opendatacommons.org/licenses/pddl/.

The Open Data Commons Attribution License is a simplified version of the ODbL. It grants the same rights, and contains most of the same restrictions, with the exception that it does contain neither the share-alike requirement nor the prohibition against including the database with technological protection measures. This makes it a very open licence, and as long as the notices are kept intact, it is very easy to comply with. It must be pointed out that both the ODbL and the Attribution licence allow commercial reuse of the database, as they both comply with the OKF's own Open Definition.[76]

The PDDL is a public domain dedication in the same spirit as CC0, but the result is a much more complex and lengthy legal document as the drafters had to contend not only with copyright, as CC0 does, but also with the database right.

This being the case, the PDDL chose to issue a dedication to the public domain of both copyright and database right similar to CC0; then it contains a waiver of those rights in case the dedication is not possible; and in case neither waiver or dedication are recognised in the local jurisdiction, then the PDDL licenses the work with a broad, unrestricted clause that reads:

> *The Licensor grants to You a worldwide, royalty-free, non-exclusive, licence to Use the Work for the duration of any applicable Copyright and Database Rights. These rights explicitly include commercial use, and do not exclude any field of endeavour. To the extent possible in the relevant jurisdiction, these rights may be exercised in all media and formats whether now known or created in the future.*

It must be said that the above makes the PDDL a very strong option for those wishing to release the work into the public domain regardless of jurisdiction.

5.3 UK Government Licensing Framework

As part of the framework arising from the PSI Directive and PSI Regulations, the UK government has been heavily involved in releasing datasets to the public by offering data through its own data portal called Data.gov.uk.[77] Parts of these efforts have been to create specific licences for public sector data.

The first licence actually dates from 2001 and it is called the Click-Use Licence, which was introduced by the Office of Public Sector Information (OPSI) in order to enable sharing of public sector information. The Click-Use licence was used particularly to enable reuse of a wide range of Crown copyright material, that is, copyright material produced by UK government departments and agencies. Similarly, the licence was used to release other public information such as laws and statutes from England and Wales.[78] However, the Click-Use approach offered only a limited solution for data, as it allowed use and reuse, but not modification. Similarly, it was also seriously under-used, as even four years after its release only approximately 7,000 licences had been issued.[79]

Given these limitations, the UK government decided to implement a new licensing scheme through the Controller of Her Majesty's Stationery Office (HMSO) called the UK Government Licensing

---

[76] The open Definition reads: "A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and/or share-alike." See: http://opendefinition.org/.

[77] http://data.gov.uk.

[78] Waelde C et al. *The Common Information Environment and Creative Commons*, Final Report to the Common Information Environment Members of a study on the applicability of Creative Commons Licences (2005).

[79] Ibid.

Framework (UKGLF). The result of their work was the creation in 2010 of two licences used to release works from the government covered by the PSI Directive: the Open Government Licence[80] and the Non-Commercial Government Licence.[81] Both of these are hosted and administered by the National Archives. It is important to point out that the perceived lack of coverage of database right within the Creative Commons framework might have prompted the government, as it prompted the Open Data Commons, to draft its own database licence instead of using CC.

Both licences are almost entirely identical. They cover both copyright and database right works, and allow the user to copy, publish, distribute, adapt and combine the information. The only difference is that, as the name suggests, the Non-Commercial Government Licence allows these reuses only if the works is not used "in any manner that is primarily intended for or directed toward commercial advantage or private monetary compensation." The Open Government Licence on the other hand allows the same reuse rights even for commercial purposes. With regards to the user's obligations, these are similar to those found in most CC licences: the user must attribute the work; must not use the information to imply official status, and must not use in a misleading manner. The government licenses go further, however, in that they forbid any use that is in breach of other local legislation including the Data Protection Act.

It is important to stress that the use of either of the two UKGLF licences is not obligatory, although their use is encouraged by the government (more about general adoption in section 5.4). Similarly, it must be pointed out that while the regime exists for the reuse of public data, it is not likely that higher education institutions will be using it to release their own databases as they are exempt from the application of the PSI Regulations. Nonetheless, we have included them in this report as the may prove as an example of a viable licence to adopt, or they could inform the drafting of terms and conditions for repositories and databases.

5.4 Licence adoption

Before looking at higher education institutional practice in detail, it is useful to know which open licences (if any) are prevalent in the wider open data scene. It is difficult at present to take a complete snapshot of licence usage and adoption, but there are some important pointers that may give an indication of the types of licences used to protect data.

The data.gov.uk repository is a good starting point because it offers daily metadata for each hosted dataset. As of 31 July 2012, the site listed 11,720 individual metadata records, of which 9,898 (84.4%) are licensed with the Open Government Licence; the rest are mostly not specified or have no licence metadata attached, and only a minority (less that 1%) use other licences. This is an impressive result, but not really surprising when one takes into account that the terms and conditions of the site clearly specify that:

> *The data and information available through www.data.gov.uk are available under terms described in the "licence" or "constraints" field of individual dataset records (meta-data). Except where otherwise noted this is the Open Government Licence.*
>
> *All dataset records (meta-data) published on www.data.gov.uk are licensed under the Open Government Licence.*

[80] Full text here: http://www.nationalarchives.gov.uk/doc/open-government-licence/.

[81] Full text here: http://www.nationalarchives.gov.uk/doc/non-commercial-government-licence/.

The above is a good indication that a clear set of licensing instructions can seriously increase specific licence adoption within an archive. Contrast that with the information gathered by a recent survey of databases in the OKF's own Data Hub catalogue.[82] This site offers no instructions, other than the fact that the site's metadata is licensed with the Open Database Licence. Of the 4,004 entries in that repository, an astounding 50% do not have any specific licence attached to them. This is surprising as the site favours open data, so one would expect a much higher level of open data sophistication. Of the datasets released with a licence, 31% used some form of CC licence, while only 11% used an Open Data Commons license. Only a minority used some form of UK government licence like the Open Government Commons (Figure 2).[83]

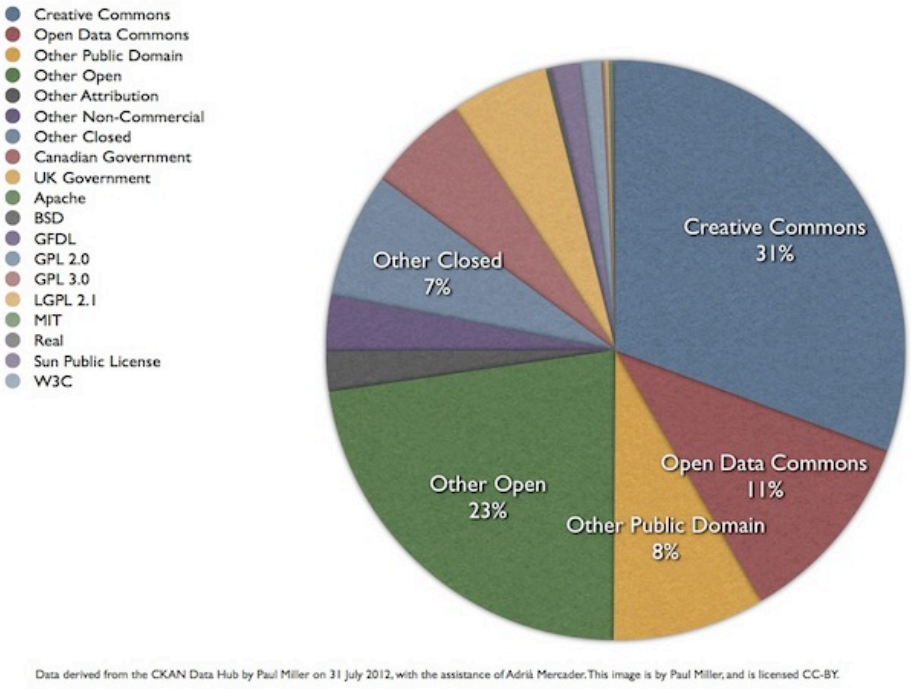

Figure 2. Types of licence used in Data Hub datasets.

This is an interesting finding for many reasons. Firstly, the current versions of Creative Commons licences are not specifically designed to work with the database right, so the datasets licensed under it may only cover the copyright element. Secondly, some of the other licences in use are not only not directed towards protecting databases, they are specifically software licences: e.g. the Apache Public License, the General Public License (GPL) in its various forms, and the Berkeley Software Distribution (BSD), just to name a few. This indicates that developers, owners and database makers in general are either not aware of other licensing choices, or they are aware of the existing licences and choose specific solutions because they are tailored to their needs. What seems clear is that licence choice is fragmented outside of the core UK government datasets, and this is not a favourable practice for potential content mining operations, as will be seen below.

---

[82] Miller P, "Thinking about Open Data, with a little help from the Data Hub", Cloud of Data (31 July, 2012), http://bit.ly/MZG5vN.

[83] Ibid.

5.5 Licence compatibility

Any attempt to measure open licence adoption may seem like an academic exercise, an attempt to distinguish between different flavours of the same thing. However, licence choice has very important consequences to reuse of content, as one licence may impose conditions that make it incompatible with other licence clauses used downstream. This is relevant particularly when dealing with collections, databases and other types of collective works. Incompatible licences could make it difficult to reuse and aggregate content from various sources.[84] This is one reason why CC licences remain very popular due to their high visibility and name recognition, as a strategist interviewed in a JISC report commented, "it's got to be CC [Creative Commons] or we're not using it. Because that just removes all the complexities."[85]

To illustrate this point, imagine a content mining project that gathers content from two different archives, one that uses a Creative Commons BY-NC-SA licence, and another one that uses the ODbL. At the time of writing, these licences are incompatible with each other because the ShareAlike element in CC licences only permits the user to distribute modified works under "the terms of this License; (ii) a later version of this License with the same License Elements as this License; (iii) a Creative Commons jurisdiction license (either this or a later license version) that contains the same License Elements as this License".[86] The ODbL contains a broader ShareAlike definition that allows the redistribution of adaptations with a "compatible license", but there is no list of compatible licences included, so in theory, both licences require derivatives to be published with their own terms. Furthermore, the NonCommercial element in the CC licence would also make it incompatible with the ODbL.

In fact, various versions within a licensing suite can be incompatible with each other. The ShareAlike and the NoDerivatives element in Creative Commons make some content released with a CC licence incompatible with other content, as Table 1 indicates.

Compatibility chart		Terms that may be used for a derivative work or adaptation						
		BY	BY-NC	BY-NC-ND	BY-NC-SA	BY-ND	BY-SA	PD
Status of original work	PD							
	BY							
	BY-NC							
	BY-NC-ND							
	BY-NC-SA							
	BY-ND							
	BY-SA							

Table 1: CC licence compatibility matrix.[87] *Green indicates compatibility.*

[84] For examples of problems with licence incompatibility in open source software, see: Rosen LE, *Open Source Licensing: Software Freedom and Intellectual Property Law*, Upper Saddle River, N.J.: Prentice Hall PTR (2004), p. 267.

[85] White D and Manton M, *Open Educational Resources: The Value of Reuse in Higher Education*, JISC Report (2011), http://bit.ly/PwT3iR.

[86] s 4 b) CC BY-NC-SA 3.0.

[87] From: http://bit.ly/TKdSud.

Creative Commons has not declared any licence as compatible at the time of writing,[88] and as we can see from the above, content released with some CC licences, such as BY-NC-ND, are incompatible with other licences for downstream reuse.

It is important to point out that some other licensing suites have been drafted to attempt to ease content interoperability; the Open Government Licence for example, has a clause that specifically covers which licences are compatible with its own terms:

> *These terms have been aligned to be interoperable with any Creative Commons Attribution Licence, which covers copyright, and Open Data Commons Attribution License, which covers database rights and applicable copyrights.*

So, content published with these terms could be remixed with content released with any of the cited legal documents. The opposite however is not always the case, so what we have is known as one-way compatibility. In practice, this means that content released with the Open Government Licence can be reused and redistributed with either CC-BY or with ODC Attribution because it clearly states that it can be done, but not the other way around. The reason for this is in the terms of various licence elements in licences such as CC. For example, the existing ShareAlike element precludes any derivatives from being shared with anything other than another CC ShareAlike licence which has the same terms and conditions. Similarly, the ND licence element precludes the creation of transformative derivatives, which would preclude content released with other licences.

The ideal situation would be to have content released with fewer licences to avoid incompatibility. This is of course, not likely given the diversity of licensing choices on display above. The other solution then is for licensors to try to maximise compatibility by trying to choose only one licence. While this is difficult, it can be done by a concerted effort from important decision makers. In the context of improving CC licence compatibility, Dulong de Rosnay suggests:

> *User communities or institutional entities (e.g., universities, Wikipedia for the BY SA 3.0, and funders) could recommend the use of only one license, as a top-down ideological prescription, after identifying the license that best suits their particular needs. For instance, in addition to making CC options' features more accessible, the CC could explain that the Share Alike clause's effect is similar to the effect of the Non-Commercial option, at least in regards to limiting commercial exploitation. The CC could also explain that reputation and integrity concerns, which often lead to the choice of the Non-Derivative options, are already ameliorated by the Attribution clause*.[89]

For the time being, potential users of incompatible content have the option of trying to gain permission to use another licence from the licensor. While this is cumbersome, it decreases legal issues arising from licence choice. It is true that many licensing institutions may not be aware of possible licence incompatibility, and may not even attempt to pursue a licence breach for the use of an incompatible licence. Nonetheless, wilful infringement is never recommended.

---

[88] See: http://creativecommons.org/compatiblelicenses.

[89] Dulong de Rosnay M, *Creative Commons Licenses Legal Pitfalls: Incompatibilities and Solutions*, IViR Report (2010), http://halshs.archives-ouvertes.fr/halshs-00671622.

6. Higher education repositories

The aforementioned strong institutional push towards open access from the UK government and important funding bodies has had a clear impact in higher education institutions. One of the most visible effects is the growth in institutional digital archive facilities, otherwise known as repositories, where academics and researchers can upload their own work in order to make it available to the public or the institution can have dedicated staff uploading, updating and maintaining such data. JISC defines digital repositories in the following manner:

> *A digital repository is a managed, persistent way of making research, learning and teaching content with continuing value discoverable and accessible. Repositories can be subject or institutional in their focus. Putting content into an institutional repository enables staff and institutions to manage and preserve it, and therefore derive maximum value from it. A repository can support research, learning, and administrative processes. They are commonly used for open access research outputs.*[90]

It is possible to classify repositories based on the type of submission. Some institutions have all-purpose repositories[91] where institutional content is stored; others have separate sites for theses, published articles and working papers,[92] while some institutions have subject specific repositories.[93]

With regards to content mining, it is important both to be able to access the contents of a repository and to have the appropriate permission to reuse the content afterwards. In this section we will analyse both.

6.1 Repository technical infrastructure

It is important to first define what is understood as a repository; technically it is not the same as a mere online collection of works. Heery enumerates the distinguishing characteristics of a true repository:

- *content is deposited in a repository, whether by the content creator, owner or third party*
- *the repository architecture manages content as well as metadata*
- *the repository offers a minimum set of basic services e.g. put, get, search, access control*
- *the repository must be sustainable and trusted, well-supported and well-managed.*[94]

---

[90] JISC, *Digital Repositories,* (2012), http://www.jisc.ac.uk/whatwedo/topics/digitalrepositories.aspx.

[91] For an example see TeesRep, the Teeside University repository: http://tees.openrepository.com/tees/.

[92] The University of Birmingham has separate sites for theses (etheses.bham.ac.uk), published articles (eprints.bham.ac.uk), and working papers (epapers.bham.ac.uk).

[93] See the Electronic Gateway for Icelandic Literature at the University of Nottingham (www.egil.nottingham.ac.uk), and the First World War Poetry Digital Archive (www.oucs.ox.ac.uk/ww1lit) at Oxford.

[94] Heery R, *Digital Repositories Review*, Report for the United Kingdom Office for Library and Information Networking (2005).

Similarly, Heery points out that most repositories follow open access principles, in which case they must provide full access to the content by members of the public, unless there are legal constraints (e.g. data protection issues), and also access to the metadata must be free. Metadata access is important for a variety of reasons, mainly because it allows cataloguing of contents, e.g. by subject, licence, institution, author, etc.

These criteria can only be met with an adequate technical infrastructure in place, preferably one that makes it easy not only to upload but also to search and access content. This is best accomplished if the information is stored with standard formats and in compliance with metadata standards.[95]

Because of the favourable policies outlined earlier, considerable investment has been made to support repository infrastructure both at the technical and logistic level. This has resulted in a technically favourable environment for content mining within the UK's higher education repositories. JISC in particular has been at the forefront of funding and supporting the development of institutional repositories. The result of such funding is a wealth of technical tools that allow ease-of-access to repository data.

Many tools have been developed to allow easier access to higher education repositories for the purpose of content mining including:

- *Directory of Open Access Repositories (OpenDOAR)*. This is a global directory of freely accessible repositories; it is operated by the SHERPA Project at the University of Nottingham. Besides linking, the directory also has a very useful tool for obtaining repository statistics.[96]
- *SHERPA Search.* This is a full-text search of all the UK repositories listed in the OpenDOAR.[97]
- *Institutional Repository Search*. One of the biggest challenges of having a vast network of institutional repositories is actually having access to to the information contained within. This project pulls content from over 130 repositories and creates a cross-search and aggregation platform.[98]
- *The RepUK Project.* This is another aggregator tool that harvests metadata from over 150 repositories. It also caches the obtained information, and offers search options by subject.[99]
- *National Centre for Text Mining (NaCTeM)*. This is a publicly funding project that offers text mining tools to academics. These include search and analysis software, training, publications and tutorials.[100]
- *JISC Standards Catalogue.* An authoritative list of recommended standards for repositories, which include everything from document standards to use of rights.

---

[95] Ibid, p.18.

[96] http://www.opendoar.org.

[97] http://www.sherpa.ac.uk/repositories/sherpasearchalluk.html.

[98] http://irs.mimas.ac.uk/demonstrator/.

[99] http://repuk.ukoln.ac.uk/.

[100] http://www.nactem.ac.uk/.

- *OpenDOAR Policy Tool*. This is an extremely useful tool for creating repository policies. This tool will be covered in more detail in the next section.[101]
- *ONIX for Publications Licenses (ONIX-PL)*. This is a family of XML formats designed to express legal terms in machine-readable form.[102]
- *JISC InfoKit on Digital Repositories*. This is a must-read for any institution setting up a repository; it contains links and explanations to everything, from software to standard.[103]

This is not an exhaustive list by any means, but it is an indication that institutional repository environment is a vibrant and dynamic field, and useful tools are constantly being produced.

6.2 Repository policies

It is evident that technical standards and tools are highly developed, but unfortunately the same cannot be said for the intellectual property issues surrounding repositories. While the open access ethos is on the rise, and the quality of content and database standard licences is also increasing, repositories do not always have clear policies on use and reuse of data and metadata. We conducted a survey of various aggregated data and of individual repositories, which produced relatively poor policy implementation.

There are several types of policies that can govern a repository. A report from the Data Information Specialists Committee-UK (DISC-UK)[104] describes the following types of policies:

- **Metadata policy:** for the information that describes items in the repository.
- **Data access and reuse policy:** for the items contained in the repository; this includes full-text works and other full data items.
- **Submission policy:** concerning various issues such as the identity of depositors, access, quality of content, formats, and most importantly for the purpose of this study, copyright policy.
- **Preservation policy:** concerns long-term issues, such as data sharing and archiving.

These four core types of policies reflect the highly complicated set of legal issues governing repositories. The IP aspects on their own are complex, as one must take into account the competing interests and needs of funders, researchers, students, and university departments. It is rare to find an institution-wide IP policy that covers all of the above parties and types of work.[105]

Researching the user terms and conditions of institutional repositories is a difficult endeavour because of the lack of clarity, and in many instances, the complete absence of policies and terms of use. We visited all of the sites linked to in the SHERPA institutional repository list,[106] looking for any

---

[101] http://www.opendoar.org/tools/en/policies.php.

[102] http://www.editeur.org/21/ONIX-PL/.

[103] http://www.jiscinfonet.ac.uk/infokits/repositories.

[104] Green A, MacDonald S and Rice R, *Policy-making for Research Data in Repositories: A Guide*, Report from the Data Information Specialists Committee-UK (2009), http://www.disc-uk.org/docs/guide.pdf.

[105] A good example of an institution that takes a holistic approach to IP is Oxford, see; http://www.admin.ox.ac.uk/rso/ip/.

[106] http://www.sherpa.ac.uk/guidance/instcontacts.html.

indication of clear terms of access and reuse. Most sites visited had a submission copyright policy in place, so the terms and conditions were centred on providing an introduction to copyright for authors. In most sites, the policies were geared towards education and avoiding the submission of papers where the author did not have copyright in the work, and therefore were designed to minimise the institution's liability.[107] This is evidenced by the presence on several sites of procedures for removal ("take down") of copyright infringing content.[108] In some instances, the absence of key policies appears to be due to the use of technology that makes it to present other documents besides the actual archive. For example, several institutions use DSpace software, which has a limited user interface that may discourage the inclusion of additional documentation.[109]

Of the 192 HEIs listed in the SHERPA institutional repository list, only 53 institutions had a publicly accessible repository, so we used those as a representative sample for analysis. Of those 53 institutional repositories visited, 45 (84%) had some sort of copyright policy, but as stated before, these were mostly for submission purposes. In fact, of the total visited, only 20 sites (37%) had clear, easy-to-access and unambiguous data reuse policies (See Apendix 1). The sample indicates that while copyright awareness is high, there is still a long way to go towards converting that awareness into reuse policies.

It must be said that, where present, many institutions offer good submission practices, attempting to ensure that the database contents themselves are not infringing copyright. The University of Leicester has a good example of a concise set of guidelines to that effect:[110]

1. *Items may only be deposited by accredited members of the institution, or their delegated agents*
2. *Authors may only submit their own work for archiving*
3. *Eligible depositors must deposit full texts of all their publications, although they may delay making them publicly visible to comply with publishers' embargos*
4. *The administrator only vets items for the eligibility of authors/depositors, and relevance to the scope of Leicester Research Archive*
5. *The validity and authenticity of the content of submissions is the sole responsibility of the depositor*
6. *Items can be deposited at any time, but will not be made publicly visible until any publishers' or funders' embargo period has expired*
7. *Any copyright violations are entirely the responsibility of the authors/depositors*
8. *If Leicester Research Archive receives proof of copyright violation, the relevant item will be removed immediately.*

Regarding submission polices, we did not find in any of the repositories any example of further granularity in the terms and conditions with regards to the origin of the work. As stated above, repositories tend to be classed as general, thesis, published article and working paper. As such, there is no indication of the source of funding, i.e. whether the funding comes from private

[107] See for example the Bristol Repository of Scholarly Eprints (ROSE), http://is.gd/1di7Rw, or the Anglia Ruskin Research Online user guide: http://libweb.anglia.ac.uk/academic/files/ARROguide.pdf.

[108] For example, see the Robert Gordon University policies: http://is.gd/og4LCA.

[109] E.g University of Hartfordshire and University of Edinburgh. The exception to this rule is the University of Leicester, which uses Dspace and has user guidelines: https://lra.le.ac.uk/.

[110] http://www2.le.ac.uk/library/about/policies/lra-policies.

enterprises, or from government sources. This lack of distinction simplifies that policy; however the same repository may include works that are subject to conflicting rights regimes.

The Directory of Open Access Repositories (OpenDOAR) contains a considerably more comprehensive list of 207 repositories in the UK. The divergence with the SHERPA list can be explained by the fact that the OpenDOAR is more updated, but that it also lists archives belonging to non-HEIs as well as various institutions that have multiple repositories (e.g. the University of Southampton hosts 11 separate ones). As stated above, many other institutions maintain separate archives for theses and for academic papers.

The OpenDOAR has conducted a thorough survey of all of the repositories listed, and its figures are similar to our sample. They look at reuse policy for both metadata and data, as many websites have different policies for each.

For metadata, 61% of UK repositories have either unknown or undefined metadata policies. Of those with one in place, 10.6% allow for commercial use, while 28.4% allow reuse only for non-profit purposes (Figure 3).

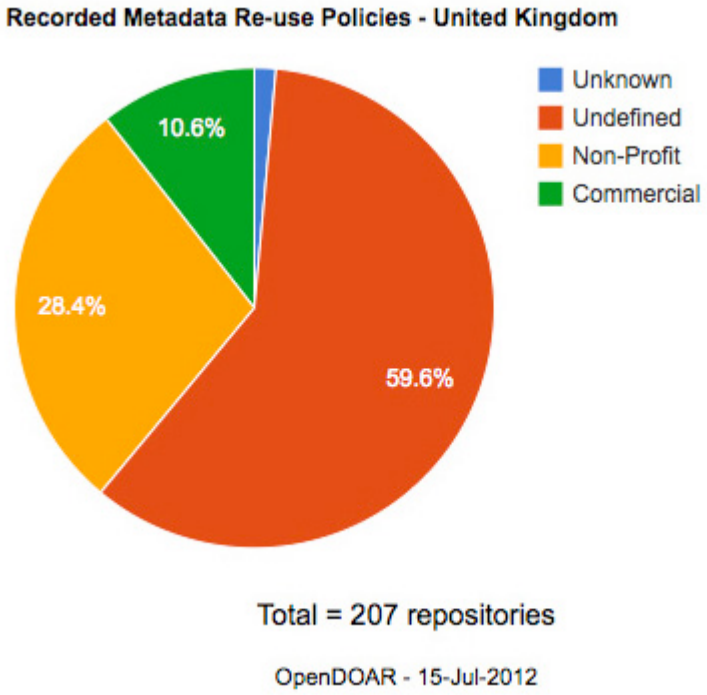

Figure 3. Recorded metadata re-use policies UK.[111]

For full-text data reuse, the OpenDOAR survey found that 57.7% of sites had an unknown, undefined or unclear policy in place. 18.8% had policies in which the rights varied for the reuse of full data items, and 13.9% only allowed reuse for non-profit purposes. Interestingly, the survey found that 8.2% of sites did not allow full-text indexing of sites by mechanical means through the existence of a No Robots file (norobots.txt). This operates as a *de facto* prohibition for reuse of data, even if such a restriction is not intended (Figure 4).

[111] http://bit.ly/N3xHMW.

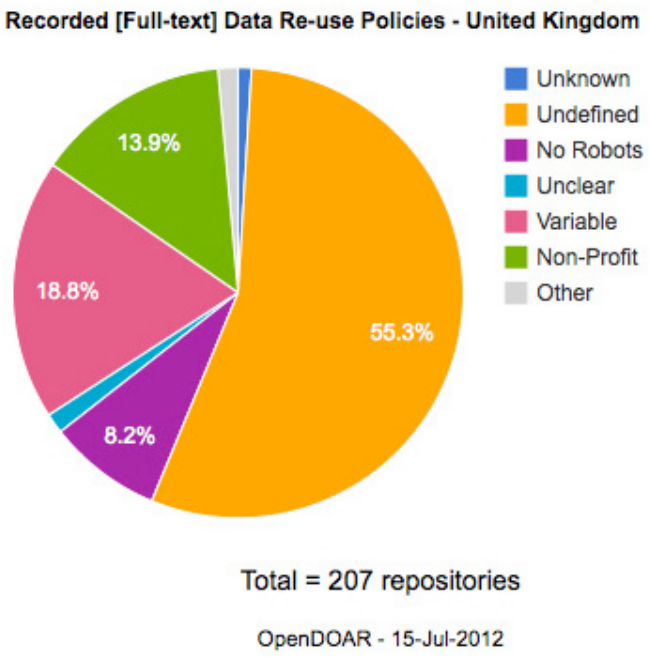

Figure 4. Recorded full-text data re-use policies UK.[112]

Both sets of statistics make for some worrying reading, as it is clear that even when available, the range of rights and restrictions on offer is too varied. When it comes to licensing, it could be said that less is more, and it would be desirable that one set of terms and conditions should prevail in one way or another, much as it does in the sample of databases licensed under the data.gov.uk site.

The source of the problem may come from the fact that these institutional repositories are not choosing their policies in a strategic manner due to the lack of harmonisation of licensing tools. Some sites are clearly using ad hoc policies,[113] while a few sites visited choose to use Creative Commons for reuse.[114] As stated before, these choices may not be compatible with databases; similarly, the reused materials from sites using CC licences incompatible with each other mean that those contents cannot be mixed without obtaining permission.

Most sites with reuse guidelines in place seem to be using the OpenDOAR Policy Tool. This is an application which generates text for five different types of policy: Metadata, Data, Content, Submission and Preservation. In each one of these fields, the institution chooses between a set of options to produce a page which can then be included in the repository. These options can be quite complex, for metadata alone users select between 10 variables, and for data there are 30 fields where selection is available. This goes a long way to explaining the statistics shown above, as it is clear that repositories are spoiled for choice. The disadvantage of this situation is that it creates

[112] http://bit.ly/N3xGID.

[113] See for example Aberystwyth University: http://bit.ly/TJM365; and the University of St. Andrews: http://bit.ly/TJMGMU.

[114] Imperial College uses the generic BY-NC-ND 3.0 http://bit.ly/N3zQrY; while the Open University uses CC BY-NC-SA 2.0 England & Wales, see: http://bit.ly/N3zDF8.

interoperability issues if one wishes to reuse data from various different datasets, as some of the elements of choice are incompatible with one another.[115]

Nonetheless, the OpenDOAR Policy Tool produces some clear policy text for both metadata and data. Take the example of the metadata policy for the Nottingham ePrints repository,[116] which is typical of many other sites:

> *Metadata Policy for information describing items in the repository*
>
> 1. *Anyone may access the metadata free of charge.*
> 2. *The metadata may be re-used in any medium without prior permission for not-for-profit purposes and re-sold commercially provided the OAI Identifier or a link to the original metadata record are given.*

Data policies generated through the tool tend to be more complex, but comprehensive. The Abertay Research Collection from the University of Abertay offers a very precise set of data access and reuse rules:[117]

> *2. Data reuse*
> *Policy for use of full-text and other full data items in the repository:*
> - *Anyone may access full items in all externally accesible Collections, apart individually embargoed items, free of charge.*
> - *Embargoed items are withheld from view due to legal requirements or to comply with publisher, funder or University policies.*
> - *Copies of open access full items generally can be reproduced, displayed or performed, and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided:*
>   - *the authors, title and full bibliographic details are given*
>   - *a hyperlink and/or URL are given for the original metadata page*
>   - *the original rights permission statement is given.*
> - *Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.*
> - *Some full items are individually tagged with different rights permissions and conditions which must be adhered to.*

Interestingly, we were not able to find a single HEI repository using either the Open Data Commons licences, or the Open Government Licence. Lack of familiarity may be to blame, or perhaps those sites that have thought about intellectual property tend to use tools that are specifically designed for repositories. Whichever reason, there is a danger of the balkanization of UK data, with government, open data, and HEI repositories all using incompatible terms and conditions.

---

[115] An example of the excessive time and cost required to secure individual permission from each source in order to aggregate their content is detailed in Box 3 High Transaction Costs in McDonald D and Kelly U, *Intelligent Digital Options and The Value and Benefits of Text Mining*, JISC report (2012), http://bit.ly/TEpc9f.

[116] Policies can be found here: http://eprints.nottingham.ac.uk/policies.html.

[117] Terms of Use can be found here: http://is.gd/LiyBoX.

6.3 Contrasting HEI policies with other repositories

While it can be said that the policy landscape in HEIs seems to be continuously improving, it may be useful to contrast it with what is taking place with other types of repositories, as well as the practices regarding content mining in the proprietary scientific publication environment.

PubMed Central UK is typical of non-HEI and non-public sector repositories in the fact that it specifies that archived works may fall under full copyright protection, and therefore cannot be considered open access. In their copyright policy, they state:

> *Articles and other material in UKPMC usually include an explicit copyright statement. In the absence of a copyright statement, users should assume that standard copyright protection applies, unless the article contains an explicit statement to the contrary. In case of doubt, contact the journal publisher to verify the copyright status of an article.*

Similarly, PubMed Central UK has strong provisions against automated and systematic download of articles:

> *Crawlers and other automated processes may NOT be used to systematically retrieve batches of articles from the UKPMC web site. Bulk downloading of articles from the main UKPMC web site, in any way, is prohibited because of copyright restrictions.*

These restrictive practices seem to be the default outside of the open access publishing community. It is calculated that in the wider PubMed Central repository, 83% of all content is not licensed to allow content mining.[118] Similarly, high-profile academics and researchers have been publicly complaining about the difficulty of accessing published works for text mining purposes,[119] which has prompted the creation of the 3 principles of open content mining mentioned above.

In an interesting project, geneticists Max Haeussler and Casey Bergman started to document their attempts to obtain permission to text mine journal articles hosted by commercial scientific publishers and their repositories.[120] This negative response from Wolters-Kluwer is typical of the replies they are getting:

> *Any reproduction, distribution, performance, display, preparation of derivative works based upon, framing, capturing, harvesting, scraping, or collection of, or creating of hypertext or other links or connections to, any Site Materials or any other proprietary information of WKH, without WKH's advance written consent, is prohibited.*

The above seems to somewhat contradict research conducted by the Publishing Research Consortium (PRC), an industry association of academic publishers.[121] In the study the authors polled 190 journal publishers. Of these, 48% said that they had detected unauthorised crawling and downloads of their content, and 51% had received requests from individual research projects. 90% of those polled claimed that they had granted access for mining for research-focused mining requests, although 69% accepted that they dealt with requests on a case-by-case basis. This means that there is no wholesale, industry-wide approach to content mining, and proprietary "all rights

---

[118] Nature "Editorial: Gold in the Text?" 483 *Nature* 124 (March 2012), http://bit.ly/Nx7c3M.

[119] Jha A, "Text mining: what do publishers have against this hi-tech research tool?" *The Guardian* (Wednesday 23 May 2012), http://bit.ly/Nx7GqD.

[120] Hosted at the UCSC Genome Bioinformatics Genocoding Project at http://text.soe.ucsc.edu/.

[121] Smit, E and Van Der Graaf M, "Journal Article Mining: The Scholarly Publishers' Perspective", 25:1 *Learned Publishing* 35 (2012).

reserved" copyright policies are the default position. There is clearly scope for improvement in this area, and this could be the subject of future studies looking in more detail at a possible change in scientific academic publishing.

7. Recommendations

Given the growing importance of content mining, it is imperative that the legal issues surrounding it should be made clear. As things stand, there are too many uncertainties in UK and this uncertainty is magnified by the fact that many databases will contain content from sources outside the UK where different rules may prevail. Content miners should be cautious and should not assume anything until they have read closely the terms and conditions governing each dataset. Government, research funding councils and HEIs all have a role to play in ensuring greater access to research for the purpose of mechanised data analysis.

1. Exceptions to copyright

Content mining does not fall easily into existing exceptions and limitations to copyright. Even when done for research purposes, the scope of fair dealing for research and personal study is too narrow. Taking that into consideration, we recommend the following points:

A. Government should push for national copyright reform that will grant an exception for text mining in accordance to the recommendation contained in the Hargreaves Review of Intellectual Property. The text in reads: "The Government should introduce a UK exception in the interim under the non-commercial research heading to allow use of analytics for non-commercial use, as in the malaria example above, as well as promoting at EU level an exception to support text mining and data analytics for commercial use."

B. This exception should be broadened as to accommodate the more generic term "content mining", as it currently reads "text mining and data analytics".

C. Include an exception for content mining for research purposes in s20 of the Copyright and Rights in Databases Regulations 1997.

D. Government should try to implement the Limits to Copyright Recommendation from the Hargreaves Review. This will make it impossible for a contract to limit exceptions and limitations to copyright law. Such a provision is needed because commercial scholarly publishers often offer conditions that either specifically or implicitly preclude content mining; much data comes from journal articles subject to subscription terms and conditions which would override the exception and negate government policy.

2. Open access

The UK is at the forefront of the enactment of public policies that favour open access. From government reports, such as the Hargreaves Review and Finch Report, to the promise to free publicly-funded works in the future, the UK public sector is taking the right steps in this regard.

A. Government should continue with its policy of promoting open access.

B. The role of JISC and other QUANGOs as facilitators to open access should continue.

C. Public funding bodies that are not already doing so should require funded research to be made accessible to the public whenever viable. These requirements should take into account the difficulty of the subject, which is reflected in the Finch Report. Funding bodies should conduct an adequate analysis of the possible effects of freeing up content as to not affect legitimate commercial interests that may hinder the UK's economy. Similarly, thought

should be given to the use of potential embargo periods to allow publishers time to recover costs.

D. Open access policies should not trump other competing interests, such as privacy and data protection. Whenever possible, a balance should be struck between the benefits of access, and the rights of data subjects.

3. Open data

There is a growing trend towards distinguishing open access with open data, with the understanding that open access pertains mostly to full-text publications, while open data deals mostly with large datasets.

A. Government and funding bodies should include specific mentions to open data into their open access recommendations, if they are not doing so already.

B. JISC, the Open Knowledge Foundation, data.gov.uk and other bodies have been pushing towards larger harmonisation of technical standards required for the viable and efficient sharing of datasets. Such technical efforts, such as the enactment of standards, the interoperability of file formats, metadata harvesting, shared search facilities, and the creation of data hubs, must continue.

4. Licensing

There is a wealth of choice of open licences which may help to enable content mining. All of the three major suites discussed in the report can prove advantageous. However, too many choices may lead to incompatibility. The case study of the UK government's data hub offers a successful example where a top-down decision pertaining licensing choices resulted in high levels of adoption of one licence.

A. The existing database licensing scene has potential compatibility issues. Whenever possible, standard licensing schemes should be encouraged.

B. Top-down recommendations from important stakeholders and repository supporting institutions may encourage licence harmonisation.

C. Database makers should consider interoperability first when choosing a licence. They should also try to choose the more free, more open and less restrictive licences, (e.g . choose BY-SA over BY-NC-SA, or choose the Open Commons Attribution Licence over the ODbL).

5. Higher education repositories

The large number, variety and scope of UK HEI repositories represent potential opportunities for content mining. In order to do so, researchers and institutions should have in place an adequate technical and legal framework that supports access.

A. HEIs should make sure that their repositories have a reuse policy in place. In the interest of content reuse compatibility, this should take the shape of an open licence (e.g. CC, ODbL, or Open Government Licence). In the absence of a licensing decision, a reuse policy should be in place (see recommended policy terms in Recommendation 6).

B. Researchers should be encouraged by the institution to make their research available by depositing it to a repository. This should be done by the implementation of an institution-wide policy that contains careful consideration of submission requirements. HEIs should

take into account existing legal requirements from commercial editors about published work.

C. Institutions should promote self-archiving of research content. This includes teaching materials, working papers, preparatory research notes, and wherever possible, published works. It should be understood that self-archiving presents difficulties for researchers, such as lack of time and insufficient technical knowledge. Whenever possible, staff should be available to support repository submission.

D. JISC, SHERPA, and other UK institutions that provide support and/or funding for HEIs should make an unequivocal choice of licence for repositories. This will help institutions make more informed decisions, will make more works available for reuse, and will also enhance interoperability of content.

E. The OpenDOAR Policy Tool is currently used by many repositories to generate their policies. This tool is valuable, but it should be overhauled to reduce the number of options available. An attempt should be made to group as many options as possible in groups of rights akin to CC's four licensing elements.

6. Standard terms and conditions

If repositories decide to choose Metadata, Data, Submission and Preservation policies, the following texts are suggested (based on the OpenDOAR Policy Tool):

Metadata Policy

A. Anyone may access the metadata free of charge.

B. The metadata may be re-used in any medium without prior permission for not-for-profit purposes and re-sold commercially provided the OAI Identifier or a link to the original metadata record are given.

Data Policy

A. Anyone may access full items free of charge.

B. Copies of full items generally can be:

- reproduced, displayed or performed, given to third parties, and stored in a database in any format or medium
- for personal research or study, educational, not-for-profit, or commercial purposes without prior permission or charge.

provided:

- the authors, title and full bibliographic details are given
- a hyperlink and/or URL are given for the original metadata page
- the original copyright statement is given
- the original rights permission statement is given
- the content is not changed in any way

C. Full items must not be sold commercially in any format or medium without formal permission of the copyright holders.

Submission Policy

A. Items may only be deposited by accredited members of the organisation, or their delegated agents.

B. The administrator only vets items for the eligibility of authors/depositors, relevance to the scope of the repository, valid layout & format, and the exclusion of spam

C. The validity and authenticity of the content of submissions is the sole responsibility of the depositor.

D. No embargo policy defined.

E. Any copyright violations are entirely the responsibility of the authors/depositors.

F. If the repository receives proof of copyright violation, the relevant item will be removed immediately.

Preservation Policy

A. Items will be retained indefinitely.

B. The repository will try to ensure continued readability and accessibility.

C. Items will be migrated to new file formats where necessary.

D. Where possible, software emulations will be provided to access un-migrated formats.

E. It may not be possible to guarantee the readability of some unusual file formats.

F. The repository regularly backs up its files according to current best practice.

G. The original bit stream is retained for all items, in addition to any upgraded formats.

H. Items may not normally be removed from the repository.

I. Acceptable reasons for withdrawal include:

J. Proven copyright violation or plagiarism

K. Legal requirements and proven violations

L. National Security

M. Falsified research

N. Withdrawn items are not deleted *per se*, but are removed from public view.

O. Withdrawn items' identifiers/URLs are retained indefinitely.

P. URLs will continue to point to 'tombstone' citations, to avoid broken links and to retain item histories.

Q. Changes to deposited items are **not** permitted.

R. *Errata* and *corrigenda* lists may be included with the original record if required.

S. If necessary, an updated version may be deposited.

T. No closure policy defined.

References

Accessibility, sustainability, excellence: how to expand access to research publications. Report of the Working Group on Expanding Access to Published Research Findings: http://is.gd/91tsKb.

Addison T et al, *The Manchester Manifesto,* Institute for Science, Ethics and Innovation (2009).

Ananiadou1 S, Kell DB, and Tsujii J, "Text Mining and its Potential Applications in Systems Biology" 24:12 *Trends in Biotechnology* 571 (2006).

Ayris P, "Why panning for gold may be detrimental to open access research", *The Guardian* (23 July 2012), http://is.gd/uscUS3.

Cannataro M, Talia D, "The knowledge grid: An Architecture for Distributed Knowledge Discovery". 46:1 *Communications of the ACM* 89 (2003).

Corley C et al, "Text and Structural Data Mining of Influenza Mentions in Web and Social Media", 7:2 *International Journal of Environmental Research and Public Health* 596 (2010).

Cornish WR and Llewelyn D, *Intellectual Property : Patents, Copyright, Trade Marks & Allied Rights*, 7th ed ed, London: Sweet & Maxwell (2010), p.509.

Davison MJ , Hugenholtz PB, "Football fixtures, horse races and spin-offs: the ECJ domesticates the database right", 3 *European Intellectual Property Review* (2005).

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society.

Dulong de Rosnay M, *Creative Commons Licenses Legal Pitfalls: Incompatibilities and Solutions*, IViR Report (2010), http://halshs.archives-ouvertes.fr/halshs-00671622.

Dusollier S, *Scoping Study on Copyright and Related Rights and the Public Domain*, Study for WIPO Committee on Development and Intellectual Property (CDIP/7/INF/2).

European Commission, *First Evaluation of Directive 96/9/EC on the Legal Protection of Databases*, DG Internal Market Working Paper, http://is.gd/DsY3XV.

Fayyad U, Piatetsky-Shapiro G, and Smyth P, "From Data Mining to Knowledge Discovery in Databases", 17:3 *AI Magazine* 37 (1996).

Fayyad U, Piatetsky-Shapiro G, and Smyth P, "From Data Mining to Knowledge Discovery in Databases", *AI Magazine* 37 (1996).

Frawley WJ, Piatetsky-Shapiro G, and Matheus CJ, "Knowledge Discovery in Databases: An Overview", 13:3 *AI Magazine* 57 (1992).

Green A, MacDonald S and Rice R, *Policy-making for Research Data in Repositories: A Guide*, Report from the Data Information Specialists Committee-UK (2009), http://www.disc-uk.org/docs/guide.pdf.

Guadamuz A, "Open Science: Open Source Licences for Scientific Research", 7(2) *North Carolina Journal of Law and Technology* 321-366 (2006).

Han J and Kamber M, *Data Mining: Concepts and Techniques*, San Francisco, CA: Morgan Kaufmann Publishers (2000).

Heery R, *Digital Repositories Review*, Report for the United Kingdom Office for Library and Information Networking (2005).

Intellectual Property Office, *Digital Opportunity: A Review of Intellectual Property and Growth*, (2011), http://www.ipo.gov.uk/ipreview.htm.

Jha A, "Text mining: what do publishers have against this hi-tech research tool?" *The Guardian* (Wednesday 23 May 2012), http://bit.ly/Nx7GqD.

JISC, *Digital Repositories,* (2012), http://www.jisc.ac.uk/whatwedo/topics/digitalrepositories.aspx.

Johnson P, "Dedicating Copyright to the Public Domain", 71:4 *Modern Law Review* 587 (2008).

Korn N, Oppenheim C and Duncan C, *IPR and Licensing issues in Derived Data*, JISC report (2007), http://bit.ly/TEmtMX.

Krallinger M, Valencia A and Hirschman L, "Linking genes to literature: text mining, information extraction, and retrieval applications for biology", 9:2 *Genome Biology* S8 (2008)

Larose DT, *Discovering Knowledge in Data: An Introduction to Data Mining*, New York, NY: John Wiley & Sons (2005).

MacQueen HL, Laurie GT and Waelde C, *Contemporary Intellectual Property: Law and Policy*, Oxford: Oxford University Press (2008), p. 66.

Madhavan M, "Copyright versus Database Right of Protection in the UK: The Bioinformatics Bone of Contention", 9:1 *Journal of World Intellectual Property* 61 (2006).

McDonald D and Kelly U, *Intelligent Digital Options and The Value and Benefits of Text Mining*, JISC report (2012), http://bit.ly/TEpc9f.

Miller P, "Thinking about Open Data, with a little help from the Data Hub", Cloud of Data (31 July, 2012), http://bit.ly/MZG5vN.

Nature "Editorial: Gold in the Text?" 483 *Nature* 124 (March 2012), http://bit.ly/Nx7c3M.

O'Connor B et al, "From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media* (2010).

OutLaw, *Database Rights: The Basics* (2008), http://www.out-law.com/page-5698.

Pang B and Lee L, "Opinion Mining and Sentiment Analysis", 2:1 *Foundations and Trends in Information Retrieval* 1 (2008).

Proudfoot RE et al, J*ISC Final Report: IncReASe (Increasing Repository Content through Automation and Services)*, White Rose Consortium (2009).

Research Information Network, *Stewardship of Digital Research Data - Principles and Guidelines*. London: RIN (2008), http://www.rin.ac.uk/data-principles.

Rosen LE, *Open Source Licensing: Software Freedom and Intellectual Property Law*, Upper Saddle River, N.J.: Prentice Hall PTR (2004).

Sample I, "Free access to British scientific research within two years", *The Guardian* (15 July 2012), http://is.gd/yOCTus.

Smit, E and Van Der Graaf M, "Journal Article Mining: The Scholarly Publishers' Perspective", 25:1 *Learned Publishing* 35 (2012).

Tan P-N, Steinbach M and Kumar V, *Introduction to Data Mining*, New York, NY: Pearson Addison-Wesley (2006).

Van Noorden R, "Trouble at the text mine", *Nature News* (7 March 2012), http://bit.ly/O78IFj.

Waelde C et al. *The Common Information Environment and Creative Commons*, Final Report to the Common Information Environment Members of a study on the applicability of Creative Commons Licences (2005).

Witten IH, Frank E and Hall M, *Data Mining: Practical machine learning tools and techniques*, New York, NY: Elsevier (2011).

Zhao K et al, "A visual data mining framework for convenient identification of useful knowledge", *Fifth IEEE International Conference on Data Mining* (2005).

Appendix

1. Breakdown of institutions with accessible policies

Institution	Repository	Reuse Policy
Aberystwyth, University of Wales	Cadair	Yes
Anglia Ruskin University	Anglia Ruskin Research Online (ARRO)	No
Aston University	Aston University Research Archive	No
Birkbeck College	Repository	No
Brunel University	Brunel University Research Archive (BURA)	No
Cardiff University	Cardiff ePrints Caerdydd	Yes
(Council for the Central Laboratory of the Research Councils)	CCLRC ePublication Archive	No
Cranfield University	Cranfield QUEprints	No
De Montfort University	De Montfort University Open Research Archive	No
Durham University	Durham Research Online	Yes
Glasgow Caledonian University	Research Online	No
Imperial College	Repository	Yes
Kings College	Repository	No
Kingston University	Repository	No
Lancaster University	Lancaster ePrints	Yes
Liverpool John Moores University	Repository	No
London School of Economics (LSE)	Repository	No
Loughborough University	Loughborough University Institutional Repository	No
Manchester Metropolitan University	e-space	No
Middlesex University	Middlesex University Digital Repository	No
Open University	Open University E-prints Service	Yes
Robert Gordon University	OpenAIR @ RGU	No
Royal Holloway	Repository	No
School of Oriental & African Studies (SOAS)	Repository	Yes
St Andrews University	St Andrews Eprints	No
The British Library	Repository	Yes
University College, London (UCL)	Repository	Yes
University of Aberdeen	Aberdeen University Research Archive (AURA)	No
University of Abertay	Abertay Research Collections (ARC)	Yes
University of Birmingham	Repository	No
University of Brighton	Repository	No
University of Bristol	Repository	No
University of Cambridge	Repository	Yes
University of Chester	ChesterRep	No
University of Edinburgh	Repository	No
University of Exeter	Exeter Research and Institutional Content archive (ERIC)	Yes
University of Glasgow	Enlighten	No

University of Hertfordshire	Repository	No
University of Leeds	Repository	No
University of Leicester	Repository	Yes
University of Newcastle	Repository	Yes
University of Nottingham	Repository	Yes
University of Oxford	Repository	No
University of Portsmouth	University of Portsmouth Eprints Archive	Yes
University of Sheffield	Repository	No
University of Southampton	e-Prints Soton	Yes
University of Stirling	University of Stirling Digital Repository	No
University of Strathclyde	University of Strathclyde Institutional Repository	Yes
University of Surrey	Repository	Yes
University of Sussex	Sussex Research Online	No
University of Wolverhampton	Repository	No
University of York	Repository	No
York St John University	Repository	Yes