Open Content Mining

Peter Murray-Rust,[*] Jennifer C. Molloy[†] and Diane Cabell[‡]

[*]University of Cambridge and OKFN, Cambridge, UK

[†]University of Oxford and Open Knowledge Foundation, Oxford, UK

[‡]Oxford e-Research Centre, Creative Commons and iCommons Ltd, Oxford, UK

Introduction

As scientists and scholars, we are both creators and users of information. Our work, however, only achieves its full value when it is shared with other researchers so as to forward the progress of science. One's data becomes exponentially more useful when combined with the data of others. Today's technology provides an unprecedented capacity for such data combination.

Researchers can now find and read papers online, rather than having to manually track down print copies. Machines (computers) can index the papers and extract the details (titles, keywords etc.) in order to alert scientists to relevant material. In addition,

How to cite this book chapter:
Murray-Rust, P., Molloy, J. C. and Cabell, D. 2014. Open Content Mining. In: Moore, S. A. (ed.) *Issues in Open Research Data*. Pp. 11–30. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/ban.b

computers can extract factual data and meaning by "mining" the content.

We illustrate the technology and importance of content-mining with 3 graphical examples which represent the state of the art today (**Figure 1–3**). These are all highly scalable (i.e. can be applied to thousands or even millions of target papers without human intervention. There are unavoidable errors for unusual documents and content and there is a trade-off between precision ("accuracy") and recall ("amount retrieved") but in many cases we and others have achieved 95% precision. The techniques are general for scholarly publications and can be applied to theses, patents and formal reports as well as articles in peer-reviewed journals.

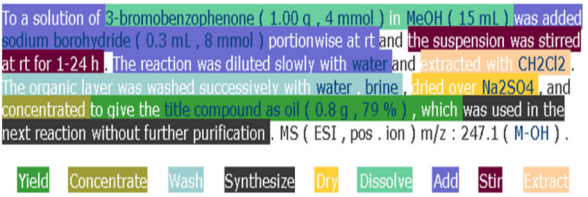
Content mining is the way that modern technology makes use of digital information. Because the scientific community is now globally connected, digitized information is being uploaded from hundreds of thousands of different sources (McDonald 2012). With current data sets measuring in terabytes, it is often no longer possible to simply read a scholarly summary in order to make scientifically significant use of such information (Panzer-Steindel & Bernd 2004; Nsf.gov, 2010; MEDLINE, 2013). A researcher must be able to copy information, recombine it with other data and otherwise "re-use" it to produce truly helpful results. Not only is mining a deductive tool to analyze research data, it is the very mechanism by which search engines operate to allow discovery of content, making connections – and even scientific discoveries – that might otherwise remain invisible to researchers. To prevent mining would force scientists into blind alleys and silos where only limited knowledge is accessible. Science does not progress if it cannot incorporate the most recent findings to move forward.

However, use of this exponentially liberating research process is blocked both by publisher-imposed restraints and by law. These

A:

To a solution of 3-bromobenzophenone (1.00 g, 4 mmol) in MeOH (15 mL) was added sodium borohydride (0.3 mL, 8 mmol) portionwise at rt and the suspension was stirred at rt for 1-24 h. The reaction was diluted slowly with water and extracted with CH2Cl2. The organic layer was washed successively with water, brine, dried over Na2SO4, and concentrated to give the title compound as oil (0.8 g, 79%), which was used in the next reaction without further purification. MS (ESI, pos. ion) m/z: 247.1 (M-OH).

B:



C:

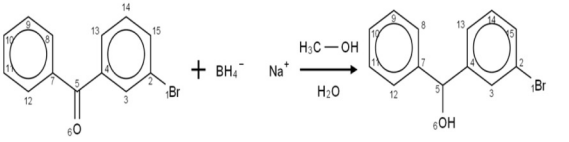
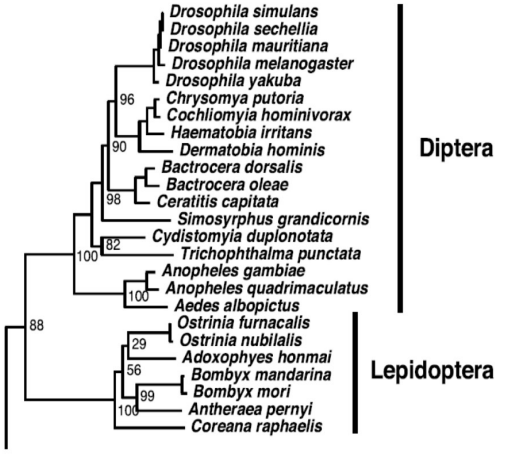


Figure 1: "Text mining". (a) the raw text as published in a scientific journal, thesis or patent. (b) Entity recognition (the compounds in the text are identified) and shallow parsing to extract the sentence structure and heuristic identification of the roles of phrases (c) complete analysis of the chemical reaction by applying heuristics to the result of (b). We have analyzed about half a million chemical reactions in US patents (with Lezan Hawizy and Daniel Lowe).

A:



B:

```
<trees label="TreesBlockFromXML" id="Trees" otus="taxl">
 <tree id="tree1" label="tree1" xsi:type="nex:FloatTree">
  <node id="N1" otu="t1" label="Psocoptera "/>
  <node id="N2" otu="t2" label="Sternorrhyncha"/>
  <node id="N3" otu="t3" label="Phthiraptera "/>
  <node id="N4" otu="t4" label="Thysanoptera "/>
  <node id="N5" otu="t5" label="Cicadomorpha "/>
  <node id="N6" otu="t6" label="Heteroptera "/>
  <node id="N7" label="N7"/>
  <node id="N8" label="N8"/>
  ...
  <edge id="line183" label="line183" source="N16" target="N1"/>
  <edge id="polyline176.3" label="polyline176.3" source="N15" target="N2"/>
  <edge id="polyline177.3" label="polyline177.3" source="N14" target="N3"/>
  <edge id="polyline178.1" label="polyline178.1" source="N11" target="N4"/>
  <edge id="polyline180.1" label="polyline180.1" source="N10" target="N5"/>
  <edge id="polyline181.1" label="polyline181.1" source="N7" target="N6"/>
  <edge id="polyline179.1" label="polyline179.1" source="N8" target="N7"/>
  <edge id="polyline175.1" label="polyline175.1" source="N9" target="N8"/>
  ...
 </tree>
</trees>
```

Figure 2: Mining content in "full-text". (a) a typical "phylogenetic tree" [snippet] representing the similarity of species (taxa) – the horizontal scale can be roughly mapped onto an evolutionary timeline; number are confidence estimates and critical for high quality work. These trees are of great value in understanding speciation and biodiversity and may require thousands of hours of computation and are frequently only published as diagrams. (b) Extraction of formal content as domain-standard (NE)XML. This allows trees from different studies to be formally compared and potentially the creation of "supertrees" which can represent the phylogenetic relation of millions of species.

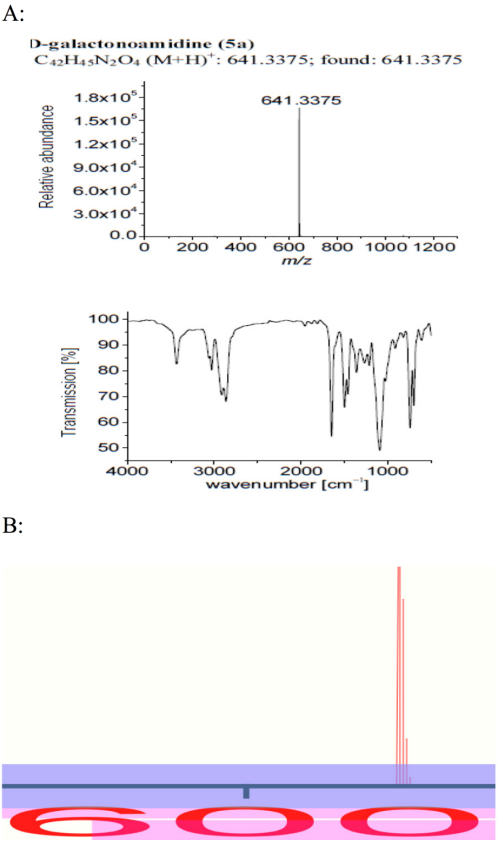


Figure 3: Content -mining from "Supplemental Data" (or "Supporting Information"). This data is often deposited alongside the "full-text" of the journal, sometimes behind the publishers firewall, sometimes openly accessible. It may run to tens or hundreds of pages and for some scientists it is the most important part of the paper. (a) exactly as published [snippet]. Note the inconvenient orientation (designed for printing) and the apparent loss of detail. (b) after content mining techniques and re-orientation – for the "m/z" spectrum (note the fine structure of the main peak, not visible in (a)). It would be technically possible to recover >> 100,000 spectra like this per year from journals.

constraints are based on business models that still rely on print revenue and are supported by copyright laws originally designed for 18th century stationers[1]. While Open Access (OA) practices are improving the ability of researchers to read papers (by removing access barriers), still only around 20% of scholarly papers are offered under OA terms (Murray-Rust 2012). The remainder are locked behind pay walls. As per the terms imposed by the vast majority of journal subscription contracts, subscribers may read pay-walled papers but they may not mine them.

Securing permission to mine on a journal-by-journal basis is extraordinarily time consuming. According to the Wellcome Trust, 87% of the material housed in UK's main medical research database (UK PubMedCentral) is unavailable for legal text and data mining (Hargreaves 2011). A recent study funded by the Joint Information Systems Committee (JISC), an association funded by UK higher education institutions, frames the scale of the problem:

In the free-to-access, UKPMC repository there are 2930 full-text articles, published since 2000, which have the word 'malaria' in the title.

Of these 1,818 (62%) are Open Access and thus suitable for text mining without having to seek permission. However, the remaining 1,112 articles (38%) are not open access, and thus permission from the rights-holder to text-mine this content must be sought.

The 1,112 articles were published in 187 different journals, published by 75 publishers.

[1] The Statute of Anne was the first UK law to provide for copyright regulation by government. See Statute of Anne, Wikipedia at http://en.wikipedia.org/wiki/Statute_of_Anne

As publisher details are not held in the UKPMC database, the permission-seeking researcher will need to make contact with every journal. Using a highly conservative estimate of one hour research per journal title (i.e., to find contact address, indicate which articles they wish to text-mine, send letters, follow-up non-responses, and record permissions etc.) this exercise will take 187 hours. Assuming that the researcher was newly qualified, earning around £30,000 pa, this single exercise would incur a cost of £3,399.

In reality however, a researcher would not limit his/her text mining analysis to articles which contained a relevant keyword in the title. Thus, if we expand this case study to find any full-text research article in UKPMC which mentions malaria (and published since 2000) the cohort increases from 2,930 to 15,757.

Of these, some 7,759 articles (49%), published in 1,024 journals, were not Open Access. Consequently, in this example, a researcher would need to contact 1,024 journals at a transaction cost (in terms of time spent) of £18,630; 62.1% of a working year (Hargreaves 2011)

Data and the Law

The intention of copyright law is to support public dissemination of original works so that the public may benefit from access to them. It accomplishes this goal by granting to authors and creators a period of monopoly control over public use of their works so that they might maximize any market benefits. While these principles may work well to protect film producers and musicians, in the current digital environment it is the unfortunate case that they actually delay or block the effective re-use of research results by the scientific community. Research scientists rarely receive any share of the profits on sales of their journal articles, but do benefit greatly by having other scientists read and cite their work. Their

interest is therefore best served by maximizing user access and use of their published results.

Databases are protected in a number of ways, most commonly by copyright and database laws. Copyright protects "creative expression" meaning the unique way that an author presents his intellectual output and it prohibits any one from copying, publicly distributing, and adapting the original without permission of the author. Specific statements of facts, shorn of any creative expression as is the case with many types of data, are themselves not ordinarily copyrightable as individual items. However, copyright does come into play for individual data points that exhibit creative expression, such as images (photographs). A collection of data can also be protected by copyright if there is sufficient creativity involved in the presentation or arrangement of the set. In the case of collections, it is only the right to utilize the collection as a whole that is restricted while the individual facts within the collection remain free.

Databases are additionally and independently protected under a *sui generis* regime imposed by the 1996 EU Database Directive (European Parliament 1996). Under the Directive, rights are granted to the one who makes a substantial investment in obtaining, verifying or presenting the contents of the database. Permission of the maker is required to extract or re-utilize all or a substantial portion of the database or to continuously extract or re-utilize insubstantial parts on a continuing basis.

To further complicate matters, copyright and database laws differ from each other and also from one jurisdiction to another. Copyrights may last for more than a hundred years (life of the author plus 70 years). Database rights (which could apply to the self-same database) only run for 15 years however those rights can be extended indefinitely by adding new data to produce a

new "work" thus triggering a new term of rights, making it horrendously difficult to determine whether or not protection has expired. The United States, for example, does not impose any *sui generis* rights. Copyright ownership belongs to the creator or his employer, but may be transferred to another (such as a publisher) hence copyright ownership can be difficult to ascertain, particularly where multiple researchers have contributed to the whole. Legal rights in such cases may be jointly held and/or held by one or more employers and/or held by one or more publishers or repositories. The authors of many "orphan" works are unknown or unidentifiable. The more globally-developed the database, the more sets of laws come into play to further complicate the definition of rights.

There are exceptions to such laws when work may be used for specific purposes without permission of the owner. In the UK, these come under the rubric "fair dealing." The UK has a current exception for noncommercial research and private study, however much research is conducted by commercial entities such as pharmaceutical companies.

Even where the law would allow free use of data, publishers imposed restrictions (**Table 1**). The terms of the user's subscription contract – deemed to be a private contract by mutually consenting parties—thus overrides any copyright or database freedoms allowed by law.

Proposed changes in legal policy

Government studies have recognized the harm such restrictions cause to the advancement of science and economic development. They argue that mining is a "non-consumptive" use that does not directly trade on the underlying creative and expressive purpose

Publisher	License Agreement Link	Explicitly prohibits text/data mining?	Quote from standard license agreement
InformaWorld	http://www.informaworld.com/smpp/termsandconditions_partiiintellectualproperty	Yes	This licence does not include any derivative use of the Site or the Materials, any collection and use of any product listings, descriptions, or prices; any downloading or copying of account information for the benefit of another merchant; or any use of data mining, robots or similar data gathering and extraction tools. In addition, you may not use meta tags or any other "hidden text" utilising our name or the name of any of our group companies without our express written consent.
Taylor Francis	http://www.tandf.co.uk/journals/pdf/terms.pdf	Yes	Incorporates Informaworld terms – see above
Elsevier/CDL	http://orpheus-1.ucsd.edu/acq/license/cdlelsevier2004.pdf	Yes	"Schedule 1.2(a) General Terms and Conditions "RESTRICTIONS ON USAGE OF THE LICENSED PRODUCTS/ INTELLECTUAL PROPERTY RIGHTS" GTC1] "Subscriber shall not use spider or web-crawling or other software programs, routines, robots or other mechanized devices to continuously and automatically search and index any content accessed online under this Agreement."

Blackwell	http://www.blackwellpublishing.com/pdf/Site_License.PDF	No	
OUP	http://www.oxfordjournals.org/help/instsitelicence.pdf	No	
Wiley	http://www.mpdl.mpg.de/nutzbed/wiley-interscience-backfile-co-nutzungsbedingung.pdf	Probably	The systematic downloading of data and the use of excerpts from databases for commercial purposes or for systematic distribution are prohibited.
ACS	http://www.mpdl.mpg.de/nutzbed/MPG_ACS_2002.pdf?la=en	Yes	Licensee (Consortium or Single Institution) acknowledges that ACS may prevent Members and their patrons, as the case may be, from using, implementing or authorizing use of any computerized or automated tool or application to search, index, test or otherwise obtain information from Licensed Materials (including without limitation any "spidering" or web crawler application) that has a detrimental impact on the use of the services under this Agreement.

(Table continued on next page)

(Table continued from previous page)

Publisher	License Agreement Link	Explicitly prohibits text/data mining?	Quote from standard license agreement
AIP	http://www.mpdl.mpg.de/nutzbed/MPG_AIP.pdf	Yes	Systematic or programmatic downloading, printing, transmitting, or copying of the Licensed Materials is prohibited. "Systematic or Programmatic" means downloading, printing, transmitting, or copying activity of which the intent or the effect is to capture, reproduce, or transfer the entire output of a journal volume, a journal issue, or a journal topical section, or sequential or cumulative search results, or collections of abstracts, articles, tables of contents. Other such systematic or programmatic use of the Licensed Materials that interferes with the access of Authorized Users or that may affect the performance of SCITATION, for example, the use of "robots" to index content, or downloading or attempting to download large amounts of material in a short period of time, is prohibited. Redistribution of the Licensed Materials, except as permitted in Section 4, without permission of the Publishers and/or payment of a royalty to the Publishers or to the appropriate Reproduction Rights Organization, is prohibited

BMJ	http://group.bmj.com/group/about/legal/bmj-group-online-licence-single-institution-licence	No	
JSTOR	http://www.jstor.org/page/info/about/policies/terms.jsp	Yes	Prohibited Uses. Institutions and users may not:... f) undertake any activity that may burden JSTOR's server(s) such as computer programs that automatically download or export Content, commonly known as web robots, spiders, crawlers, wanderers or accelerators;
Nature	http://www.nature.com/libraries/site_licenses/2010acad_row.pdf	Yes	3. USAGE RESTRICTIONS Except as expressly permitted in Clause 2.1, the Licensee warrants that it will not, nor will it licence or permit others to, directly or indirectly, without the Licensor's prior written consent: (j) make mass, automated or systematic extractions from or hard copy storage of the Licenced Material.

Table 1: Publisher content mining policies.

of the original work or compete with its normal exploitation. Most recently, the 2011 Government-sponsored Hargreaves Report on intellectual property reform, found:

Researchers want to use every technological tool available, and they want to develop new ones. However, the law can block valuable new technologies, like text and data mining, simply because those technologies were not imagined when the law was formed. In teaching, the greatly expanded scope of what is possible is often unnecessarily limited by uncertainty about what is legal. Many university academics – along with teachers elsewhere in the education sector – are uncertain what copyright permits for themselves and their students. Administrators spend substantial sums of public money to entitle academics and research students to access works which have often been produced at public expense by academics and research students in the first place. Even where there are copyright exceptions established by law, administrators are often forced to prevent staff and students exercising them, because of restrictive contracts. Senior figures and institutions in the university sector have told the Review of the urgent need reform copyright to realise opportunities, and to make it clear what researchers and educators are allowed to do. (Hargreaves 2011)

Hargreaves recommended that the Government 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. It argues that it is "not persuaded that restricting this transformative use of copyright material is necessary or in the UK's overall economic interest." (Hargreaves 2011)

Hargreaves also urged the government to change the law at both the national and EU level to prevent any copyright exceptions from being overridden by contract.

Applying contracts in that way means a rights holder can rewrite the limits the law has set on the extent of the right conferred by copyright. It creates the risk that should Government decide that UK law will permit private copying or text mining, these permissions could be denied by contract. Where an institution has different contracts with a number of providers, many of the contracts overriding exceptions in different areas, it becomes very difficult to give clear guidance to users on what they are permitted. Often the result will be that, for legal certainty, the institution will restrict access to the most restrictive set of terms, significantly reducing the provisions for use established by law. Even if unused, the possibility of contractual override is harmful because it replaces clarity ("I have the right to make a private copy") with uncertainty ("I must check my licence to confirm that I have the right to make a private copy"). The Government should change the law to make it clear no exception to copyright can be overridden by contract" (Hargreaves 2011)

The current U.K. government also believes that the ability for research to power economic development will be greatly enhanced if content mining is encouraged. In responding to Hargreaves, the Government stated its intention to:

- bring forward proposals for a substantial opening up of the UK's copyright exceptions regime, including a wide non-commercial research exception covering text and data mining, and
- aim to secure further flexibilities at EU level that enable greater adaptability to new technologies, and

- make the removal of EU level barriers to innovative and valuable technologies a priority to be pursued through all appropriate mechanisms. (HM Government 2011)

Further, the Government believes that it is not appropriate for "certain activities of public benefit such as medical research obtained through text mining to be in effect subject to veto by the owners of copyrights in the reports of such research, where access to the reports was obtained lawfully." (HM Government 2011)

Because science is a global enterprise, change in copyright law at the national and regional levels will not be sufficient to allow the free flow of information throughout the scientific community. Such changes must be made at many national and regional levels if the goal of a free and open exchange of data is to be achieved.

Changes in publication policies

Because publishers can override legal freedoms by enforcing restrictive terms of use in subscription agreements, we urge researchers to not only support these Government initiatives, but to go further by taking personal and institutional responsibility for establishing open mining practices in their work and publishing environments. In particular, we urge the adoption of the following Open Mining Manifesto (Murray-Rust 2012).

Open Mining Manifesto

1. Define 'open content mining' in a broad and useful manner

'Open Content Mining' means the unrestricted right of subscribers to extract, process and republish content manually or by machine in whatever form (text, diagrams, images, data, audio,

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

[1] Text
[2] Numbers
[3] Tables: numerical representations of a fact
[4] Diagrams (line drawings, graphs, spectra, networks, etc.): Graphical representations of relationships between variables, are images and therefore may not be, when considered as a collective entity, data. However, the individual data points underlying a graph, similar to tables, should be.
[5] Images and video (mainly photographic)- where it is the means of expressing a fact.
[6] Audio: same as images – where it expresses the factual representation of the research.
[7] XML: Extensible Markup Language (XML) defines rules for encoding documents in a format that is both human-readable and machine-readable."
[8] Core bibliographic data: described as "data which is necessary to identify and / or discover a publication" and defined under the Open Bibliography Principles [15].
[9] Resource Description Framework (RDF): information about content, such as authors, licensing information and the unique identifier for the article.

2. Urge publishers and institutional repositories to adhere to the following principles:

Principle 1: Right of Legitimate Accessors to Mine

We assert that there is no legal, ethical or moral reason to refuse to allow legitimate accessors of research content (OA or otherwise)

to use machines to analyse the published output of the research community. Researchers expect to access and process the full content of the research literature with their computer programs and should be able to use their machines as they use their eyes. **The right to read is the right to mine**

Principle 2: Lightweight Processing Terms and Conditions

Mining by legitimate subscribers should not be prohibited by contractual or other legal barriers. Publishers should add clarifying language in subscription agreements that content is available for information mining by download or by remote access. Where access is through researcher-provided tools, no further cost should be required. **Users and providers should encourage machine processing**

Principle 3: Use

Researchers can and will publish facts and excerpts which they discover by reading and processing documents. They expect to disseminate and aggregate statistical results as facts and context text as fair use excerpts, openly and with no restrictions other than attribution. Publisher efforts to claim rights in the results of mining further retard the advancement of science by making those results less available to the research community; such claims should be prohibited. **Facts don't belong to anyone.**

3. Strategies

Assert the above rights by:

- Educating researchers and librarians about the potential of content mining and the current impediments to

doing so, including alerting librarians to the need not to cede any of the above rights when signing contracts with publishers
- Compiling a list of publishers and indicating what rights they currently permit, in order to highlight the gap between the rights here being asserted and what is currently possible
- Urging governments and funders to promote and aid the enjoyment of the above rights.

Editor's note

This article originally was originally presented at the Conference for the Fellows of OpenForum Academy, 24th September 2012 in Brussels, and is reproduced in accordance with the CC BY licence and with kind permission of the authors. Whilst there have been no alterations to the content, the reference style has been amended for consistency with the other chapters in the book.

References

European Parliament (1996) *Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases*. Available at at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31996L0009:EN:HTML [accessed 22 Sep. 2014].

Hargreaves (2011) *Digital Opportunity, A Review of Intellectual Property and Growth* available at http://www.ipo.gov.uk/ipreview-finalreport.pdf [accessed 22 Sep. 2014].

HM Government (2011) *The Government Response to the Hargreaves Review of Intellectual Property and Growth*. Available at http://www.ipo.gov.uk/ipresponse-full.pdf.

Panzer-Steindel, Bernd (2004) *Sizing and Costing of the CERN T0 center*, CERN-LCG-PEB-2004-21 available at: http://lcg-computing-fabric.web.cern.ch/lcg-computing-fabric/presentations/Sizing%20and%20costing%20of%20the%20CERN%20T0%20center.doc [accessed 22 Sep 2014].

McDonald, (2012) The Value and Benefits of Text Mining, Section 3.3.8, JISC Report Doc #811, available at http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx.

MEDLINE (2014). *MEDLINE® Citation Counts by Year of Publication (as of mid – November 2013)*. [online] Available at: http://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html [Accessed 22 Sep. 2014].

Murray-Rust P (2012) *The Right to Read is the Right to Mine* http://blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine/ [accessed 22 Sep. 2014].

Nsf.gov, (2010). *nsf.gov – Science and Engineering Indicators 2010 – Chapter 5. Academic Research and Development – Highlights – US National Science Foundation (NSF)*. [online] Available at: http://www.nsf.gov/statistics/seind10/c5/c5h.htm [Accessed 22 Sep. 2014].

Open Knowledge Foundation (2011) *Principles on Open Bibliographic Data*. Available at http://openbiblio.net/principles/ [Accessed 22 Sep. 2014].