Digital Transformation and the Changing Role of News Media in the 21st Century, August 13-14, 2014, International Telecommunication Union, Geneva, Switzerland
Submitted on: 30 June, 2014

Corpus Protocols: digital transformations of commercial newspaper collections for text and data mining to support academic research

Seth Cayley; Mike Gardner; Katherine Gupta; Michaela Mahlberg; Neil Smyth; Stella Wisdom

Abstract

This paper reports on outcomes from the Corpus Protocols project that investigated opportunities and challenges for data and text mining commercial news content, including licensing, storage and accessibility of data, and communication between researchers and publishers. Related research in digital humanities and digital libraries has concentrated on methodologies, overviews and accessibility. This paper focuses on research in corpus linguistics based on digital news data, including the Declassified Documents Reference System and the Times Digital Archive. Modernising Copyright and published draft regulations for UK legislation indicate new opportunities for libraries to support research using digital newspapers. Some publishers have pledged to bring more content online, and others are exploring sustainable commercial models for widening access to big data. This paper explores issues of research data management, innovations in web technologies for big data, and how research based on this kind of data satisfies requirements for the security of commercial news data in the context of emerging legislation. We identify potential conflicts this raises for research libraries and researchers.

Introduction

This paper reports on outcomes from the *Corpus Protocols*[1] project that investigated opportunities and challenges for data and text mining commercial text archives held by university libraries. Questions that the project dealt with include: licensing; storage and accessibility of data; and, communication between researchers and publishers. The paper focuses on research in corpus linguistics and uses the *Declassified Documents Reference System (DDRS)* as its main example. It will also point to implications for other data, including news data bases, such as the *Times Digital Archive*. The paper explores in particular how research based on this kind of data satisfies requirements for the security of commercial text archives in the context of emerging United Kingdom (UK) legislation for research, education and libraries, including text and data mining.

Digital humanities and the tensions between research methods and legal frameworks

Corpus Protocols was not a research project that started with a research question originating solely in an academic department. The aim of the project was to explore interdependencies between

[1] *Corpus Protocols* was funded by Horizon Digital Economy Research (http://www.horizon.ac.uk), and is part of the *Data-Asset-Method* network (http://www.nottingham.ac.uk/humanities/digital/dam.aspx).

academic research, interests of publishers who provide data services, the legal frameworks that shape research, as well as university infrastructures that enable research. With the increasing availability of digital data an understanding of such interdependencies is crucial for innovative research to take place. Research in digital humanities and digital libraries related to text and data mining news and other types of social sciences data has concentrated on methodologies, overviews and accessibility. The IFLA World Library and Information Congress 2013 included an overview of humanities and social sciences research methods that involve the mining of newspapers (Cheney, 2013), an assessment of the challenges for librarians (Okerson, 2013), and IFLA international newspaper conferences have focused on models for improving accessibility (eg. Allen, 2010). The IFLA Statement on Text and Data Mining states:

> that it is committed to the principle of freedom of access to information, and the belief that information should be utilised without restriction in ways vital to the educational and cultural well-being of communities, IFLA believes TDM to be an essential tool to the advancement of learning, and new forms of creation
>
> (IFLA, 2013: online)

In linguistics, especially in corpus linguistics, researchers are interested in large computer-readable data sets such as newspaper archives. Corpus linguistics investigates language on the basis of such large collections of data to find evidence for the way in which linguistic phenomena are used. Corpus linguistics uses quantitative methods that are related to techniques in data mining, but also employs qualitative methods to analyse data in its situational, social and cultural contexts. There are examples of large scale studies using newspaper data for corpus linguistics (Gabrielatos, McEnery, Diggle, & Baker, 2012; Mahlberg & O'Donnell, 2008; O'Donnell, Scott, Mahlberg, & Hoey, 2012; Partington, 2010), or research combining approaches in linguistics and sociology (eg. Grundmann & Scott, 2012), or historical research projects with innovative text mining methods (eg. Humanities in the European Research Area, 2014). One of the contexts for current and future research funding is the recognition of the potential growth and value of big data (eg. McKinsey Global Institute, May 2011), and how the big data revolution has been identified by the UK government as one of eight great technologies (Willetts, 2013).

Proposed changes to UK copyright legislation have provided one context for the project. Some of the text and data mining techniques that have been used in non-commercial academic research, including corpus linguistics, involve copying works, such as newspapers, for analysis. There have been risks of copyright infringement unless specific permissions were obtained from the rights holder. Some research studies have attempted to articulate the different issues around content mining: the UK higher education context (eg. Guadamuz, 2014; Guadamuz & Cabell, 2013); the potential changes to copyright legislation (eg. Hellwig, 2013); and wider technological developments in the European context (eg. European Commission, 2014). The Hargreaves Review illustrated how UK copyright legislation is falling behind what is needed to support research, particularly in relation to new technologies that enable text and data mining (Hargreaves, 2011). The UK Government response to the Hargreaves Review accepted the proposals including “a wide non-commercial research exception covering text and data mining” where access had been obtained lawfully (HM Government, 2011). This was followed by a Consultation on changes to copyright in the United Kingdom, including proposals for data analytics for non-commercial research that would be balanced with protecting publishers from large-scale copyright infringement (HM Government, 2012: online).

The UK government has now made important changes to copyright law for the digital age, which have important implications for text and data mining news and newspaper collections. *The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014* came into force on 1 June 2014, introducing a new section 29a that states:

> the making of a copy of a work by a person who has lawful access to the work does not infringe copyright in the work provided that (a) the copy is made in order that a person who has lawful access to the work may carry out a computational analysis of anything recorded in the work for the sole purpose of research for a non-commercial purpose, and (b) the copy is accompanied by a sufficient acknowledgement.

(*The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014. No. 1372*, 2014: online).

The Intellectual Property Office has made an *Illustrative draft of the Copyright, Designs and Patents Act* available so that the changes can be assessed in the full context of the existing UK legislation (Intellectual Property Office, 2014c). Nevertheless, the *Explanatory Memorandum* makes it clear that "this new copyright exception would permit UK researchers carrying out non-commercial research to use text and data mining technologies without risking copyright infringement" (Intellectual Property Office, 2014b: online). Moreover, the Intellectual Property Office has made further guidance available that is specific to research, emphasising what is now possible without infringing copyright (Intellectual Property Office, 2014a). A letter clarified the UK government's position that the law does not apply only to new contracts; the change applies to all contracts from the date of legislation (House of Lords Secondary Legislation Scrutiny Committee, 2014b). This change to the legal framework in the UK creates exciting opportunities and possibilities for using commercial news data.

Some publishing organisations have expressed concerns about changes to the legislation (eg. John Wiley & Sons, 2012; Veerasingham, 2012). Some publishers have pledged to bring more content online (eg. European Commission, 2013). Others are exploring sustainable commercial models for widening access to big data (eg. Smit, van de Graaf, & Publishing Research Consortium, 2011) or developing policies and technical solutions for text and data mining in the new UK legal context, such as Elsevier (Elsevier, 2014; Hersh, 2014). The changes to copyright legislation in the UK mean that University library collections and publisher data are more valuable to research because there is more that can be done with the collections, content and data. There are opportunities to create new partnerships that will link open data and publisher discovery tools, as several projects have explored (eg. Somerville & Conrad, 2014). In particular, there are new opportunities to use the news and newspaper data in existing databases where a University already has a right of access for text and data mining content.

Corpus Protocols

The project title 'Corpus Protocols' reflected our of aim of exploring different perspectives on digital assets and the need to find 'protocols' for the way in which researchers, publishers, libraries and wider university support structures negotiate the requirements and practicalities of an adequate research context that enables innovative research. Researchers in corpus linguistics would ideally like to have access to raw data so that they can access the data with corpus linguistic software or add linguistic annotation to the data for more advanced searches. O'Donnell et al. (2012), for example, investigated whether words have preferences to occur in certain positions in newspaper articles, such as in text-initial sentences. Their research could not have been easily completed by accessing newspapers through a data base like *Nexis*, but required access to raw Guardian newspaper data. The *CLiC*[2] tool developed at the University of Nottingham also shows that specific research questions need access to data in a certain way. *CLiC* makes it possible, for instance, to search for patterns in fictional speech. These patterns can only be retrieved because the novels that are accessed through the interface have been marked-up accordingly.

[2] CLiC is available at the University of Nottingham: http://clic.nottingham.ac.uk:8080/index.html

At the start of the project we knew about existing University owned data sets, such as the *Times Digital Archive* where the University had a right to access through subscription. We also knew about, or we were involved in using and developing, existing research tools and research methods for corpus linguistics. And, we knew about draft legislation for data analytics that might provide new opportunities for researchers and libraries to leverage investments in digital library resources that have been provided by publishers, such as Cengage Learning. However, we did not know the best way to make data available for researchers or the level of demand for data related services. Should this data be on the researcher computer, in a University owned data centre or accessed through an external commercial cloud? To what extent can existing research tools, such as Wordsmith[3], be used to handle large commercial data sets? To what extent are services provided by publishers informed by current and emerging research methods? *Corpus Protocols* aimed to prototype an approach for using locally stored data using existing corpus linguistic tools. Focusing on textual data (a 'corpus') we reviewed relationships between institutional infra-structure, external business partners and research questions and methods that deal with textual data, exploring opportunities for the future through the development of protocols. An existing corpus linguist tool was used to analyse a University owned data set that was made available on the University network.

The Declassified Documents Reference System as a Corpus

Corpus Protocols started with opportunistic data that was available in the library. Libraries have lawful access to many copyright protected works, including commercial newspaper collections, either through subscriptions or payment for perpetual access to the data. The University of Nottingham, for example, has perpetual access to four ProQuest historical newspaper collections (*Guardian and Observer, Los Angeles Times, New York Times* and the *Washington Post*) and we subscribe to the *Times Digital Archive* provided by Cengage Learning. In addition, we have access to many other news and newspaper sources, including *19th Century British Library Newspapers*, the 17th-18th century *Burney Collection* of newspapers, *China Daily*, *Chronicling America: Historic American Newspapers*, *Current Digest of the Post-Soviet Press*, the *Eighteenth Century Journals Portal* and *Nexis*. Library users normally access the content of these databases through publisher provided platforms and keyword searches, and this approach satisfies many research requirements.

Although the library had access to many newspaper databases, the news data was not available for the project. However the aims of the project were about exploring issues around storage, accessibility, licencing and the protocols between publishing and universities. One data set that was available and is related to newspaper sources was the *DDRS*. The *DDRS* is one of several valuable sources for research based on declassified United States of America documents. Other important sources include: *Foreign Relations of the United States,* with volumes from 1861-1960 being available for free online; the *National Security Archive*, available for purchase as the *Digital National Security Archive*; and leaked documents on web sites, such as *WikiLeaks (Gibbs)*. There are many research projects related to 20th century history, the United States of America foreign relations, national security, international relations, politics and government, nuclear strategy, the Vietnam War or the Cold War that have used *DDRS* (eg. Brands, 2012; Cooley & Spruyt, 2009; Cullather, 2011; Gavin, 2012; Holden, 2004; Khalidi, 2009; Miller, 2013; Schoultz, 2009; Wang, 2008; Zimmermann, 2001). *After Hiroshima : the United States, race, and nuclear weapons in Asia, 1945*-1965 by Matthew Jones is an example of an American foreign relations research project that used the *DDRS* including information cables and many memoranda of meetings (Jones, 2010). Figure 1 shows the

[3] Wordsmith is a program for analysing texts and corpora developed by Mike Scott: http://www.lexically.net/wordsmith/

results of a basic keyword search for the word *nuclear* which would be how most researchers have traditionally used the *DDRS*.

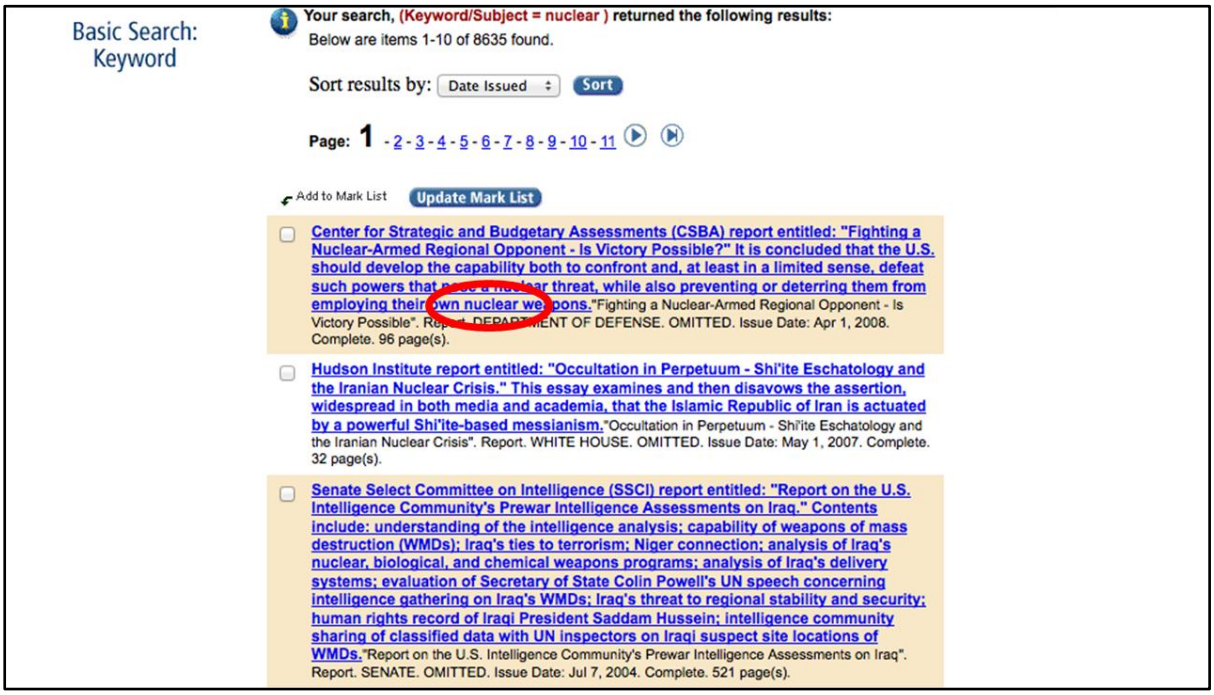

Figure 1: results from *DDRS* based on a basic keyword search for the word "nuclear".

When the University of Nottingham library invested in the *DDRS*, Cengage Learning supplied the data on a hard drive to be stored by the library. The library did not acquire and store the data thinking that it would be used for research. There was never any intention that the data would be used for research projects that used text and data mining methods. Indeed, Cengage Learning supplied the drive purely for the purposes of providing a backup to its customers in case the product became permanently unavailable. However, the publisher no longer sends such drives to customers by default, storing its products with Portico (www.portico.org) instead, and charging a fee to those customers who do request a hard drive backup (Cengage Learning). The *Corpus Protocols* project investigated the opportunities and challenges for data and text mining commercial content, including licensing; storage and accessibility of data; and, communication between researchers and publishers.

For this project we wanted to investigate how we could make this data available on the University network for analysis. There is an existing University of Nottingham procedure and infrastructure for academic researchers to request space for research projects. During the *Corpus Protocols* project requesting space for 40GB was straight forward, and 673,060 *DDRS* files (39GB) were copied to the University network. This is a different way to access the *DDRS* data from the normal way using the publisher database. In this protocol the researcher browses to the University network, selects the relevant folder for the project, and is then able to view the files containing the *DDRS* data. More importantly the data is available on the University network for text and data mining.

However, the project discovered technical problems with this approach. Firstly, for data security access was restricted to specific people, in this case the members of the project team, because anybody with access to the data can copy all the files. Moreover, extending access to all project team members was time consuming, as not a default step in the university system, indicating that there may be problems extending access to data to other groups, such as postgraduate research students. Providing access to data from non-University computers, such as personal laptops or a

home computer, was complex. University computers are mapped to the relevant University network, but for some researchers other options need to be considered, including: a web interface, which might work for one file but not for running a specific corpus linguist tool, such as Wordsmith; Citrix virtual desktop, which would work to view some files, but you could not install programs like Wordsmith for text analysis; and remote desktop to a University computer, which would work if you have a University computer but the machine would need to be left switched on or woken remotely. Web based solutions could not be delivered during the *Corpus Protocols* project, but this is an area that requires further investigation, especially for new tools that balance accessibility and security, and which reassure publishers who are providing the content.

Even in the context of changing UK legislation, publishers will have high expectations for the security of data. The Explanatory Memorandum accompanying the new UK legislation says that the exception is not a right to mine works where the researcher does not have a right to access and that publishers can "impose reasonable measures to maintain stability and security of their computer networks" (Intellectual Property Office, 2014b: online). From a publisher's perspective, any copies of commercial works made for the purposes of non-commercial mining will need to be strictly managed by the body which has purchased the access rights, which in higher education is usually the university library. Publishers will need to be reassured that the potential for further copying and dissemination that infringes copyright is minimal. The perceived risk is that once a rogue copy of the data is more widely available for sharing, it is virtually impossible to retrieve the various copies, and the damage to the publisher's commercial interests is done. This is no small matter; the data is a key part of the value that publishers create, especially with products containing digitised primary sources. A newspaper from 1785 will in its original physical form be out of copyright, but the XML data that was created at huge expense in the 21st century for the digital archive of that newspaper is very much the publisher's copyright, and a key commercial asset. With this in mind, publishers will want copies of data created for the purposes of mining to be held on a highly secured network. Access should be monitored and restricted to known users, who can be held accountable in the event of infringement. Putting such procedures in place is in a University's interests also, as the UK copyright legislation is very clear that full legal consequences will apply to any institution's users that copy and share data beyond the provisions of the law.

The University of Nottingham has an aspirational *Research Data Management Policy (University of Nottingham, 2014)*, and the related *Information Security Policy* is built into the University's management of risk at the highest level (University of Nottingham, 2013). Such security procedures admittedly reduce flexibility and restrict access to data within a university. These measures may make it more difficult to extend the mining of data sets beyond, 'just' researchers, to the classroom. Publishers will inevitably be wary about students having access to copies of data, as today's students are regarded as a group with relatively poor understanding and regard for copyright law. However, Cengage Learning feels that the benefits of text and data mining research may come down to students through new tools and forms of data access that can be developed through collaboration with researchers. Such tools would therefore provide controlled access to data without risking security and widespread copyright infringement. During the course of the project the University policies were shared in confidence with Cengage Learning. Following the changes to the UK legislation there will be opportunities for improved licencing agreements that address concerns around access to data.

In addition, publishers need to build any new business models for data services that support the University research community on a shared understanding of research trends, methods and outputs. One of the key problems for publishers, and to some extent those working in professional services at universities, is understanding the demand for and value of data, including any text and data mining approaches to using the data. In this case the focus was corpus linguistics.

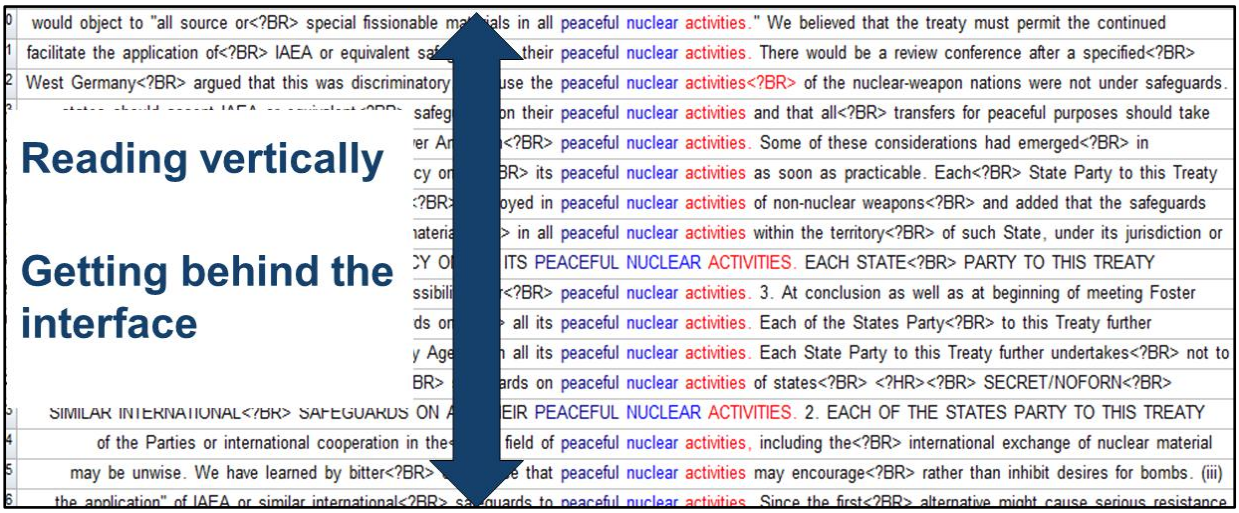

Figure 2: the keyword "nuclear" in context.

Figure 2 shows a screen shot of a concordance, i.e. display format that shows a word, here the example *nucleus* in its context. Running concordances is a basic research method for corpus linguistics that allows the researcher to identify patterns that would not become apparent through the standard database interface for the *DDRS*. In the example above, *nuclear* collocates with *activities* on the right and *peaceful* on the left.

Another advantage of being able to access the data behind the commercial interface is to organise the data into different subsets. The *DDRS* contains 197,837,328 words in total. This 'corpus' can be devided into 30 yearly sub-corpora of between 3.3 and 8.3 million words each. These subcorpora can then be used to generate 'key words', i.e. words that occur significantly more frequently in one subcorpus than in the rest of the corpus. The lists of key words in Figure 3 have been generated by comparing a year (1996, 1997 etc.) against the rest of the *DDRS*. Key words can be the starting point for further analysis, such as to find out how key words characterise specific discourses. Such statistically significant 'key words' are quite different from 'key words', in the sense of 'search words' as illustrated in Figure 1 above.

1996	1997	1998	1999	2000	2001	2002	2003	
	NBSP	ILLEGIBLE	LIBRARY	ILLEGIBLE	ILLEGIBLE	ILLEGIBLE	ILLEGIBLE	ILLEG
IOTOCOPY	BR	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT	TEXT
SSEMINATION	DR	DWIGHT	ILLEGIBLE	WE	PHOTOCOPIED	WE	ARCHIVES	NARA
J	PHOTOCOPY	LIBRARY	WE	TO	TO	I	REPRODUCED	EO
RALD	CARTER	EISENHOWER	CARTER	CARTER	WE	E	NARA	SENSI
FUGEES	TOP	TO	DWIGHT	LIBRARY	EISENHOWER	T	AUTHORITY	AUTH
RD	YOU	T	LBJ	THAT	THE	LIBRARY	DATE	AND
E	THIEU	E	COPY	LBJ	OUR	CARTER	I	TO
AT	GERALD	CHINESE	TO	WOULD	NUCLEAR	COPY	NATIONAL	THE
ASSIFICATION	FORD	LBJ	EISENHOWER	OUR	LIBRARY	LBJ	O	YOU
	WE	D	FORD	REGARDING	THAT	YOU	DECLASSIFIED	ISRAE
RARY	BRZEZINSKI	THAT	SECRET	COPY	WOULD	TO	AT	A
NSITIVE	SECRET	THE	NODIS	THE	BE	#	THE	KISSI
NTROLS	GDS	R	TEL	WITH	WITH	R	YOU	WE
TICLE	MISSILE	HE	THAT	SOVIETS	IN	KISSINGER	NIXON	IS
OULD	THINK	COPY	OUR	AND	S	NARA	EO	ISSUE
	TEST	LAO	WOULD	SAID	DWIGHT	REPRODUCED	KISSINGER	IN
UCLEAR	S	WE	NUCLEAR	BE	RELATIONS	THAT	E	I
AS	U	S	GERALD	U	U	EXDIS	RICHARD	NLS
BA	MR	SAID	NOT	HE	PAKISTAN	O	TO	ARE
IEU	BUNKER	C	YOU	NUCLEAR	REGARDING	STATE	A	ISRAE

Figure 3: *nuclear* in sub-corpora structured according to time

The type of corpus research illustrated raises questions for those developing data based services. What types of projects emerge because of access to data and larger volumes of data? Who owns the intellectual property rights in any non-standard research outputs that emerge from projects that use text and data mining techniques? Are there any areas where there is a lack of clarity where protocols are needed between publishing and non-commercial research. Are any new research outputs owned by the researcher or the publisher? Any new business models will also need to be built on an understanding of the research outputs for text and data mining projects.

***TIDAL*: the Times data and other news data for future research**

In collaboration with Cengage Learning, *TIDAL*, the *TImes Data Archive Lab* project, will investigate social reality in the 19th century based on news discourse to complement the picture of Dickensian London. This research project will be based on the *Times Digital Archive* newspaper corpus, building on the knowledge and understanding gained through *Corpus Protocols*. The project will explore new technical options, including the processing of data using High Performance Computing.

Beyond *TIDAL*, there will be opportunities to use British Library news data. The British Library has one of the world's greatest news archives. The collection of UK, Irish and world newspapers numbers over 60 million issues, from the 17th century to the present day, and there are growing collections of television, radio and web news.

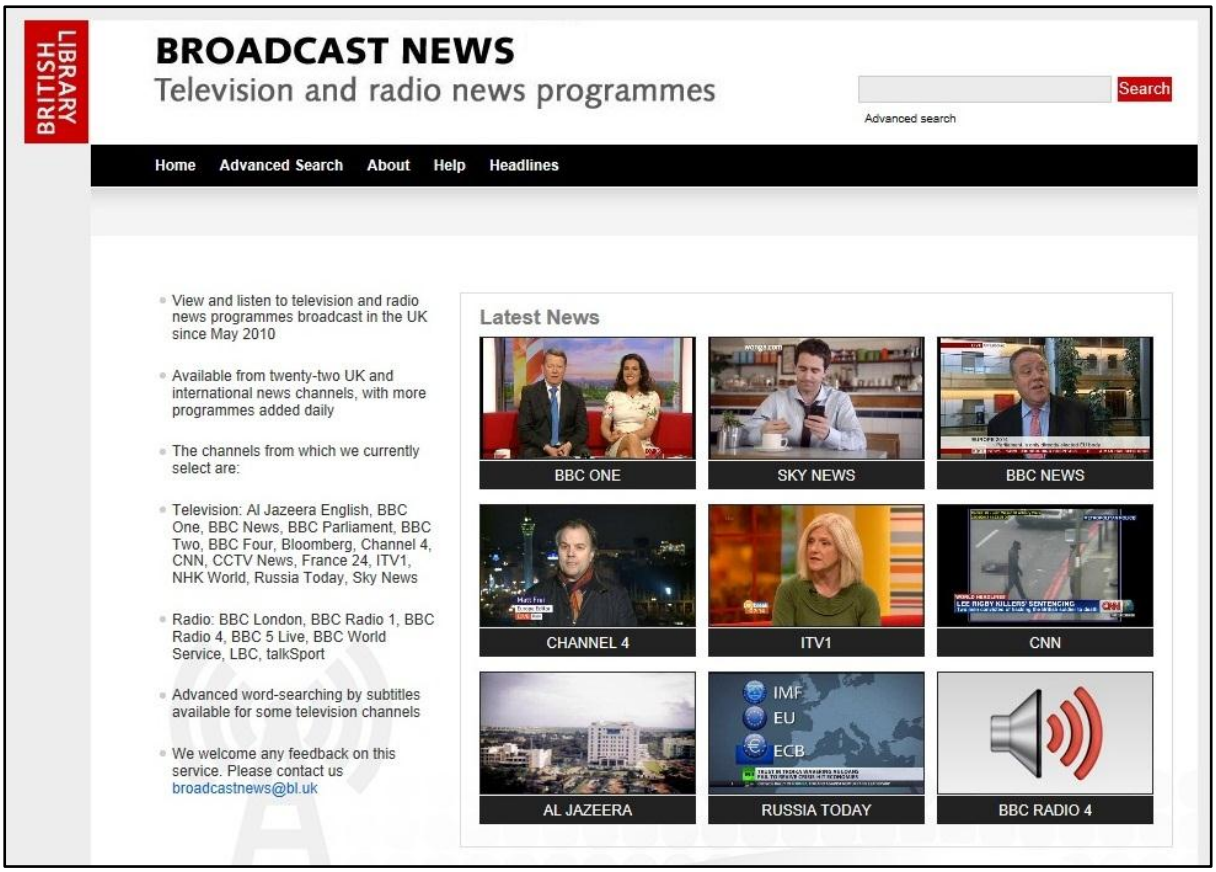

Figure 4: Broadcast News: Television and radio news programmes

Through the *Broadcast News* service daily television and radio news programmes broadcast in the UK since May 2010 are available through an instant access service in the Reading Rooms of the British Library (Hulme). Over 60 hours of news are recorded every day from 22 channels, including the BBC, ITV, Channel 4, Sky News, France 24, Bloomberg, Russian Today, China's CCTV News, Al Jazeera English and CNN. Many of the recorded *Broadcast News* television and radio news programmes come with subtitles, which are a searchable research resource. Like with the *DDRS* and the *Times Digital Archive*, these subtitles are a corpus of data and metadata that has potential value for research. Again, in the context of the new UK legislation, there will be opportunities to combine news corpora, including television news and newspaper data in new research projects that create new corpora and new non-standard research outputs.

Text and data mining newspaper content continues to be an emerging area of research. It has been recognised, for example, that the UK Regulations have the potential to be both positive for the research sector and negative for rights holders (House of Lords Secondary Legislation Scrutiny Committee, 2014a). There will be a review by the Intellectual Property Office before April 2019. Moreover, harmonisation across boarders will continue to be an issue for collaborative non-commercial research in universities (Universities UK & UK Higher Education International Unit, 2014). There are opportunities for universities and publishers to work together in partnership to assess the impact of legislation. Digital humanities research projects also have the knowledge, skills and technologies that can transform business by helping to shape the further development of newspaper database and news data services.

Conclusion

This paper explored issues of research data management, innovations in web technologies for big data, and how research based on this kind of data satisfies requirements for the security of commercial news data in the context of developing UK legislation. Changes to legislation offer new opportunities for researchers, university libraries and publishers, but there are many challenges as larger volumes of commercial data become available. The *Corpus Protocols* project illustrated corpus methods to publishers, librarians and web technologists, who all have crucial roles in supporting research. There were several important outcomes from the project: dialogue with Cengage Learning raised awareness of specific security requirements for data; the University library ensured the research process was compliant with proposed, emerging and new legislation; and technology and network security teams provided the information systems infrastructure to make commercial data available on a University network. While this might seem obvious in hindsight, the project highlighted the protocols of communication and governance for research using digital news data. Changes to legislation offer new opportunities for researchers, university libraries and publishers, but there are many challenges as larger volumes of commercial data become available.

References

Allen, R.B. (2010). *Improving access to digitized historical newspapers with text mining, coordinated models, and formative user interface design.* Paper presented at the IFLA International Newspaper Conference: Digital Preservation and Access to News and Views, New Delhi. Retrieved from: http://boballen.info/RBA/PAPERS/IFLA2010/iflaDelhi.pdf. [Retrieved: 8 February 2014].

Brands, Hal. (2012). *Latin America's Cold War*. Cumberland, RI, USA: Harvard University Press.

Cengage Learning. *Gale Extends Partnership with Portico to Preserve Remaining Digital Collections* Retrieved from: http://news.cengage.com/library-research/gale-extends-partnership-with-portico-to-preserve-remaining-digital-collections/. [Retrieved: 19 June 2014].

Cheney, Debora. (2013). *Text mining newspapers and news content: new trends and research methodologies* Retrieved from: http://library.ifla.org/233/1/153-cheney-en.pdf. [Retrieved: 4 January 2014].

Cooley, Alexander, & Spruyt, Hendrik. (2009). *Contracting States : Sovereign Transfers in International Relations*. Princeton, NJ, USA: Princeton University Press.

The Copyright and Rights in Performances (Research, Education, Libraries and Archives) Regulations 2014. No. 1372. (2014). London: The Stationery Office Limited. Retrieved from: http://www.legislation.gov.uk/uksi/2014/1372/pdfs/uksi_20141372_en.pdf. [Retrieved: 9 June 2014].

Cullather, Nick. (2011). *Hungry World : Americas Cold War Battle Against Poverty in Asia*. Cambridge, MA, USA: Harvard University Press.

Elsevier. (2014). *http://www.elsevier.com/about/universal-access/content-mining-policies* Retrieved from: http://www.elsevier.com/about/universal-access/content-mining-policies. [Retrieved: 7 March 2014].

European Commission. (2013). *Licences for Europe: ten pledges to bring more content online* Retrieved from: http://ec.europa.eu/internal_market/copyright/docs/licences-for-europe/131113_ten-pledges_en.pdf. [Retrieved: 4 January 2014].

European Commission. (2014). *Standardisation in the area of innovation and technological development, notably in the field of text and data mining. Report from the Expert Group*. Brussels: European Commission. Retrieved [Retrieved: 2 June 2014].

Gabrielatos, Costas, McEnery, Tony, Diggle, Peter J., & Baker, Paul (2012). The peaks and troughs of corpus-based contextual analysis. International Journal of Corpus Linguistics. *International Journal of Corpus Linguistics, 17*(2), 151-175.

Gavin, Francis J. (2012). *Nuclear Statecraft : History and Strategy in America's Atomic Age*. Ithaca, NY, USA: Cornell University Press.

Gibbs, David N. *Guide to Using Declassified Documents*. Arizona: University of Arizona. Retrieved from: http://dgibbs.faculty.arizona.edu/guide_using_declassified_documents. [Retrieved: 3 March 2014].

Grundmann, Reiner, & Scott, Mike. (2012). Disputed climate science in the media: Do countries matter? *Public Understanding of Science, 0*(0), 1–16. doi: 10.1177/0963662512467732

Guadamuz, Andres. (2014). Data mining in UK higher education institutions: law and policy. *Queen Mary Journal of Intellectual Property, 4*(1), 3-29. doi: 10.4337/qmjip.2014.01.01

Guadamuz, Andres, & Cabell, Diane. (2013). *Analysis of UK/EU law on data mining in higher education institutions*. Retrieved, from http://www.technollama.co.uk/wp-content/uploads/2013/04/Data-Mining-Paper.pdf.

Hargreaves, Ian. (2011). *Digital Opportunity: A Review of Intellectual Property and Growth*: Intellectual Property Office. Retrieved from: http://www.ipo.gov.uk/ipreview-finalreport.pdf. [Retrieved: 3 July 2013].

Hellwig, Frank. (2013). *Change in copyright law as market intervention to realize the welfare potential of text mining scientific research*. Leipzig, Germany: Leipzig U niversity of Applied Sciences. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2386238. [Retrieved:

Hersh, Gemma. (2014). *How does Elzevier's text mining policy work with new UK TDM law? Posted on 8 June 2014* Retrieved from: http://www.elsevier.com/connect/how-does-elseviers-text-mining-policy-work-with-new-uk-tdm-law. [Retrieved: 19 June 2014].

HM Government. (2011). *The Government Response to the Hargreaves Review of Intellectual Property and Growth*. Newport, United Kingdom: Intellectual Property Office. Retrieved from: http://www.ipo.gov.uk/ipresponse-full.pdf. [Retrieved: 26 July 2013].

HM Government. (2012, December 2012). *Modernising Copyright: a modern, robust and flexible framework; Government response to consultation on copyright exceptions and clarifying copyright law*. Retrieved, from http://www.ipo.gov.uk/response-2011-copyright-final.pdf.

Holden, Robert H. (2004). *Armies Without Nations : Public Violence and State Formation in Central America, 1821-1960*. Cary, NC, USA: Oxford University Press.

House of Lords Secondary Legislation Scrutiny Committee. (2014a). *41st Report of Session 2013-14. House of Lords Paper 180*. London: The Stationary Office Limited. Retrieved from: http://www.publications.parliament.uk/pa/ld201314/ldselect/ldsecleg/180/180.pdf. [Retrieved: 10 June 2014].

House of Lords Secondary Legislation Scrutiny Committee. (2014b). Work of the Committee in Session 2013-14 (Subsidiary House of Lords Secondary Legislation Scrutiny Committee, Trans.). In Secondary House of Lords Secondary Legislation Scrutiny Committee (Ed.), Secondary Work of the Committee in Session 2013-14 (pp. 29-30). London: The Stationary Office Limited. (Reprinted from: Reprint Retrieved from: http://www.publications.parliament.uk/pa/ld201314/ldselect/ldsecleg/186/186.pdf. [Retrieved:

Hulme, Tom. *Broadcast News at the British Library: Critical Discourse Analysis (CDA)* Retrieved from: http://www.esrc.ac.uk/_images/Broadcast-News-at-the-British-Library_tcm8-23401.pdf. [Retrieved: 2 June 2014].

IFLA. (2013). *IFLA Statement on Text and Data Mining* Retrieved from: http://www.ifla.org/files/assets/clm/statements/iflastatement_on_text_and_data_mining.pdf. [Retrieved: 5 February 2014].

Intellectual Property Office. (2014a). *Exceptions to copyright: Research*. Newport: Intellectual Property Office. Retrieved from: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/315014/copyright-guidance-research.pdf. [Retrieved: 9 June 2014].

Intellectual Property Office. (2014b). *Explanatory Memorandum* Retrieved from: http://www.legislation.gov.uk/ukdsi/2014/9780111112717/pdfs/ukdsiem_9780111112717_en.pdf. [Retrieved: 9 June 2014].

Intellectual Property Office. (2014c). *Illustrative draft of the Copyright, Designs and Patents Act* Retrieved from: http://www.ipo.gov.uk/cdpa1988-unofficial.pdf. [Retrieved: 9 June 2014].

John Wiley & Sons. (2012). *John Wiley & Son*. Newport, United Kingdom: Intellectual Property Office. Retrieved from: http://www.ipo.gov.uk/response-2011-copyright-jwsons.pdf. [Retrieved: 23 July 2013].

Jones, Matthew. (2010). *After Hiroshima : the United States, race, and nuclear weapons in Asia, 1945-1965*. Cambridge: Cambridge University Press.

Khalidi, Rashid. (2009). *Sowing Crisis : The Cold War and American Hegemony in the Middle East*. Boston, MA, USA: Beacon Press.

Mahlberg, M., & O'Donnell, M. B. (2008). *A Fresh View of the Structure of Hard News Stories.* Paper presented at the Online Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop, Saarbrücken, 23–25 July 2007. Retrieved from: http://scidok.sulb.uni-saarland.de/volltexte/2008/1700/pdf/Mahlberg_ODonnell_form.pdf. [Retrieved: 9 May 2013].

Miller, Edward. (2013). *Misalliance : Ngo Dinh Diem, the United States, and the Fate of South Vietnam*. Cumberland, RI, USA: Harvard University Press.

O'Donnell, M. B., Scott, M., Mahlberg, M., & Hoey, M. (2012). Exploring text-initial words, clusters and concgrams in a newspaper corpus. *Corpus Linguistics and Linguistic Theory, 8*(1), 73-101. doi: 10.1515/cllt-2012-0004

Okerson, Ann. (2013). *Text & Data Mining - A Librarian Overview* Retrieved from: http://library.ifla.org/252/1/165-okerson-en.pdf. [Retrieved: 3 February 2014].

Partington, Alan. (2010). Modern Diachronic Corpus-Assisted Discourse Studies (MD-CADS) on UK newspapers: an overview of the project. *Corpora, 5*(2), 83-108.

Schoultz, Lars. (2009). *That Infernal Little Cuban Republic : The United States and the Cuban Revolution*. Chapel Hill, NC, USA University of North Carolina Press.

Smit, Eefke, van de Graaf, Maurits, & Publishing Research Consortium. (2011). *Journal article mining: a research study into Practices, Policies, Plans…..and Promises. Commissioned by the Publishing Research Consortium* Retrieved from: http://www.publishingresearch.org.uk/documents/PRCSmitJAMreport2.30June13.pdf. [Retrieved: 9 September 2013].

Somerville, Mary M., & Conrad, Lettie Y. (2014). *Collaborative Improvements in the Discoverability of Scholarly Content: Accomplishments, Aspirations, and Opportunities. A SAGE White Paper*. Los Angeles, CA: SAGE. Retrieved from:

http://www.sagepub.com/repository/binaries/pdf/improvementsindiscoverability.pdf. [Retrieved: 6 June 2014]. doi: 10.4135/wp140116.

Universities UK, & UK Higher Education International Unit. (2014). *Response to the public consultation on the review of the EU copyright rules conducted by the European Commission, Directorate General Internal Market and Services. March 2014* Retrieved from: http://www.international.ac.uk/media/2562241/uuk-and-iu-response-to-ec-consultation-on-eu-copyright.pdf. [Retrieved: 3 June 2014].

University of Nottingham. (2013). *Information Security Policy 2012/13* Retrieved from: http://workspace.nottingham.ac.uk/download/attachments/62358464/IS+Security+Policy.pdf [University of Nottingham access only] [Retrieved: 9 June 2014].

University of Nottingham. (2014). *Research Data Management: policies* Retrieved from: http://www.nottingham.ac.uk/research/research-data-management/creating-data/policies.aspx. [Retrieved: 9 June 2014].

Veerasingham, Daisy. (2012). *Consultation on proposals to change the UK's copyright system (Reference 2011-004). Response submitted by Associated Press - 21 March 2012*. Newport, United Kingdom: Intellectual Property Office. Retrieved from: http://www.ipo.gov.uk/response-2011-copyright-press.pdf. [Retrieved: 23 July 2013].

Wang, Zuoyue. (2008). *In Sputnik's Shadow : The President's Science Advisory Committee and Cold War America*. New Brunswick, NJ, USA: Rutgers University Press.

Zimmermann, Hubert. (2001). *Money and Security : Troops and Monetary Policy in Germany's Relations to the United States and the United Kingdom*. Port Chester, NY, USA: Cambridge University Press.