Submitted on: 30/07/2013

Archiving Social Media in the Context of Non-print Legal Deposit

Helen Hockx-Yu
Head of Web Archiving, British Library, London, United Kingdom.
E-mail address: Helen.hockx-yu@bl.uk

Abstract:

Social media is the collective name given to Internet-based or mobile applications which allow users to form online networks or communities based on common interest, social or ideological orientations. Such applications take many forms but their main purpose is to support interaction and communication among the members of a community, including the creation and exchange of user-generated content. Twitter, Facebook and YouTube are examples of large social networking platforms, which aggregate many forms of media into one place and are used globally for business, research and personal communications. Social media have become increasingly prevalent in people's lives and also important sources for scholars to understand our time.

Social media have also become an important area of consideration and a key challenge for web archiving institutions. This is perhaps the most demanded content by researchers, yet a combination of legal, curatorial and technical issues has made archiving of social media content a non-trivial task. To date there are no scalable solutions to preserving such content. There is thus a need to address the problem collectively, starting with discussing common issues and practices.

This paper provides an overview of the current approaches to archiving social media including their pro and cons. It also reports on how the British Library addresses this, discussing the key considerations and decisions both prior to and after the non-print Legal Deposit regulations became effective – these were introduced by the UK government in April 2013. It will in addition examine the initiatives and possibilities outside the web archiving community, offering thoughts on new models potentially appropriate to archiving social media content.

This paper intends to provoke thoughts about some fundamental questions, e.g. whether our current approaches are valid, and whether national libraries should prioritise archiving public social media content.

Keywords: Web Archiving, Social Media, Non-print Legal Deposit.

1. Key challenges

Social media is the collective name given to Internet-based or mobile applications which allow users to form online networks or communities based on common interest, social or ideological orientations. Such applications take many forms but their main purpose is to support interaction and communication among the members of a community, including the creation and exchange of user-generated content. Twitter, Facebook and YouTube are key examples of large social networking platforms, which aggregate many forms of media into one place and are used globally for business, research and personal communications. Social media have become increasingly prevalent in people's lives and also important sources for scholars to understand our time.

For memory institutions which archive and preserve content on the Internet including the World Wide Web, social media have become an important area of consideration. Its pervasiveness and increasing importance to scholarly research require institutions to make conscious decisions with regard to the collection and access of a separate class of digital content which is significantly more complex.

Archiving social media inherits all the exiting legal, curatorial and technical challenges related to web archiving, documented comprehensively by the recent ISO Technical Report (Technical Committee ISO/TC 46, 2012).[1] There are however additional challenges which make archiving of social media content an even more involved task.

Websites and social media content are copyrighted and archiving them without permissions breaches the copyright law. Unless exempted by legislative frameworks for the purpose of collecting national heritage or public records, many archiving institutions have taken a permission-based approach to address copyright issues. IRP owners are informed of the intention and purpose of archiving, either in an opt-in or opt-out fashion. When applying this to social media, where intellectual property rights apply to both user-generated content and the social networking platforms that deliver it, the opaqueness of ownership, often involving multiple organisations and individuals, makes it extremely difficult to identify IPR owners and to obtain clearance for archiving. A known social network response to a permission request was that from Facebook to the UK Parliamentary Archive, on 10 June 2009, stating that "a formal permission is not required" for the archive to crawl their own Facebook pages, implying Facebook's general policy.

Boyd and Ellison (Boyd D; Ellison N, 2007) provide an early account of scholarship based on social networks, emerging from diverse disciplinary and methodological traditions, including a section on privacy issues. These concern potential threats to privacy and security and are still valid and equally applicable to accessing archived social media content. Analysis of aggregated social media datasets may reveal unexpected or unintended patterns and connections, raising concerns about the risk of breaching personal privacy. There is also a requirement on archiving institutions to provide maximum levels of transparency in the applications used to collect and analyse social media data as hidden algorithms would only deepen the concern about privacy.

[1] A formally published, nearly identical and paid-for version of the report can be found at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=55211

One of the curatorial challenges in web archiving has to do with boundary or scope. Elements of the web are intrinsically linked to each other through hyperlinks. Unless one intends to archive the global web, boundaries need to be put in place to limit the scope by instructing the web crawler to stop collecting content at some point. It is more complicated to scope social media content for archiving due to the highly participatory, interactive and instant nature. Pennock (Pennock, 2013) refers to archiving Twitter as archiving "conversations" and poses a set of interesting questions for curatorial consideration.

Web crawlers are commonly used to capture snapshots of websites. It works well with static content which is explicitly referenced in HTML text. The crawler simply downloads a copy of the targeted file via a HTTP request. The problem is that an increasing amount of the content on the web is generated dynamically and displayed through the use of client side scripts, from specific queries to a site's database or based on some user interactions. The current crawling technology is inadequate in dealing with such dynamic content, which one would encounter extensively on social media platforms. Driven by the desire to offer advanced and cutting-edge user experience, complex technologies are used to publish content and provide functionalities, and refreshed much more frequently than for example personal or organisational websites. Additional barriers are often in place to restrict the amount and frequency of the data that they let out to robots or crawlers. Pop et. al (Pop R; Vasile G; Masanes J, 2010) provide a detailed description of the long, tedious and complex process one has to go through to unravel and discover the direct URL that would enable a web crawler to reach the desired video on YouTube.

2. Current approaches

It is not surprising that memory institutions are yet to develop common and scalable solutions for archiving systematically and providing access to social media content. Effort in this to date is mostly experimental, with the exception of the Library of Congress and Twitter collaboration described below, and the current approaches can be divided into a few broad categories.

<u>Collaborative agreement between publisher and memory institution</u>

The most significant and effective archiving effort for social media so far is the agreement between the Library of Congress and Twitter, signed in in April 2010, giving the library copies of public tweets dating back to the company's inception in 2006, an archive of 21 billion tweets. The two organisations additionally agreed that public tweets will be donated to the Library on an on-going basis beyond April 2010. As of 1 December 2012 the Library of Congress has received 170 billion tweets totalling over 130TB of data (Library of Congress, 2013).

Both organisations should be applauded for this tremendous endeavour to collect and preserve the "story of America". It is hard to imagine any other alternative way which would allow more comprehensive and systematic collection of an entire social network of such scale. It is hoped that other key social media service providers would follow the Twitter model and put in place long term sustainable archival arrangements.

Understandably the Library of Congress has not yet offered research access to the Twitter Archive. Dealing with an unprecedented digital collection of massive scale, velocity, variety and complexity, it would require years of research and development until technically viable

and cost-effective access solutions can be found. It is expected that the Library of Congress would focus on acquiring and organising the Twitter Archive initially and the Archive will remain "dark" for some time.

Use and repurpose existing web archiving technology

Among memory institutions, the most common approach to archiving social media content is a selective one based on the use of existing web archiving tools. This only captures a small number of social media pages, for the purpose of specific topic collections or capturing the social network presence of individuals and organisations. It suffers from all the limitations described in the previous section. A certain level of engineering effort is often required to customise the crawling process or modify the tools. There are a small number of Twitter and Facebook pages and YouTube videos scattered in various web archives which are publically available. There should be more archived social media content within large national web archives but these often have restricted access so difficult to quantify.

Social network pages in web archives are often incomplete, due to the limitations of the crawler in dealing with dynamic content. However, the intention of archiving is to capture and replay the content, the look and feel and the context as much as possible. Despite the missing elements, these are captured to some extent. Some archives have started to compensate for the quality issues by taking full-size screenshots of the pages using a dedicated browser. When embedded and replayed as parts of a web archive, it is not always easy to find social network content, unless these are highlighted in some way or adequate search functions are in place to help users find the relevant material.

Some institutions employ external service providers who are more advanced or specialise in social media archiving. Archive-It (https://www.archive-it.org/) is an example of such a service. Searches of "Twitter" and "Facebook" in Archive-It lead to a large number of archived collections and pages. Commercial service providers in general are able to focus on technical development based on customers' demand. Archive-It, for example, issues best practice and specific instructions for their customers on how to scope social media sites for archiving. (Archive-It, 2014) They are also developing Umbra, to be used in combination with the commonly used crawler Heritrix, to improve the capture of dynamic web content. (Archive-It, 2014).

Collecting data through APIs

The UK National Archives (TNA) recently launched the UK Government Social Media Archive, beginning to archive systematically tweets and YouTube videos published by UK central government departments from their official Twitter and YouTube social media platforms. (The National Archives, 2014).

The TNA worked with their service provider the Internet Memory Foundation (IMF) to develop capturing and replay solutions for the social media archive, which make use of social media service providers' Application Programming Interfaces (APIs). This is an example of another category of approach to archiving social media, which ensures better quality, more capture (within terms and conditions defined by the service provider) and offers flexibility and possibility for customisation.

Both Twitter and YouTube offer a variety of APIs, allowing the retrieval of raw tweets and videos with associated metadata. This is the preferred way for social media service providers to make (some) data available for reuse and analytics. Twitter for example limit the number of requests to the server per hour and have a set of rules associated with the use of the APIs. It is important to realise that the APIs are primarily means to promote products and services and the terms and conditions can change at the service provider's discretion. There may be some vulnerability to rely on the APIs and collect data for long-term archiving purposes.

Further processing and a user interface are required to make the raw social media data accessible. The screenshot below from TNA's Archive shows the archived Twitter page of the UK Department for Culture, Media and Sport, which echoes the Twitter interface. It is clearly not a reproduction of the original page but a recreation instead, which could raise questions with regard to authenticity. The same however could be said about the crawled or harvested copies in web archives which often miss elements of what was once online due to technical limitations. The question is how much "look and feel" contributes to the authenticity of digital resources and where one draws the line.

TNA's Twitter Archive includes the tweets and some of the linked resources, which are also downloadable in JSON or XML. It does not aggregate tweets belonging to a conversation nor include accounts mentioned in the direct tweets, missing some of the context in which the tweets were created or placed.

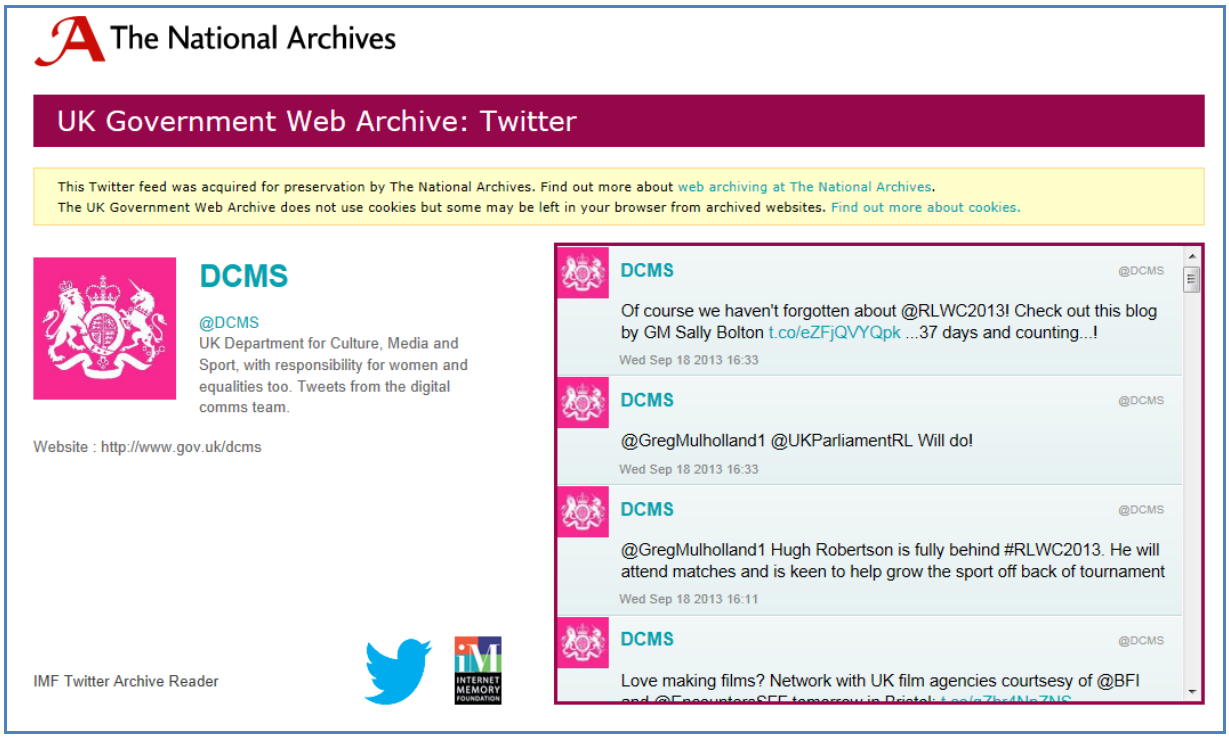

Figure 1. Archived DCMS Tweets from the UK National Archives' Twitter Archive

The US-based social media archiving service provider ArchiveSocial (http://archivesocial.com/) has taken a similar approach to collecting raw data but goes much further in order to archive and represent an organisation's social network presence and the wider context. It aggregates content from multiple accounts across a range of social networks and has a faceted search interface to allow users to explore various dimensions of archived

content. The screenshot below is the social media archive of Snohomish County, Washington, powered by ArchiveSocial, which covers a large numbers of accounts and major social media platforms. The interface allows navigation per account and/or per social network. It captures the original content as well as any replies or posts abut that content. In the case of Twitter, it includes all tweets, mentions, favourites and even direct messages. For each entry, it also displays the original timestamp and the source data, which can be downloaded (Snohomish County, Washington - Social Media Archive, 2014).

This could be a source for debate but in the context of authenticity it may be a better option not to mimic the original social network platform if the user interface is entirely recreated, to avoid confusion or misinterpretation.

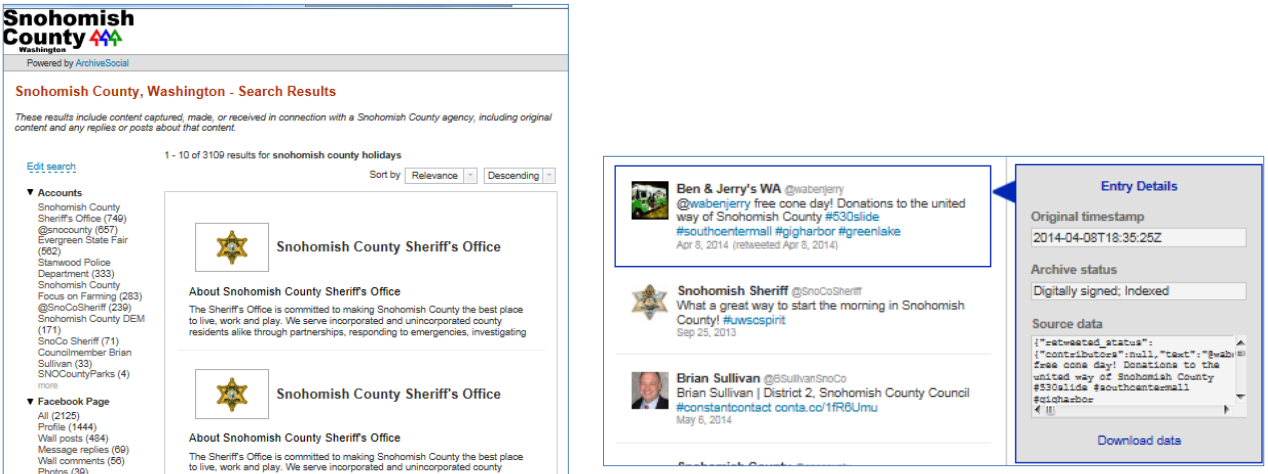

Figure 2. Social media archive of Snohomish County, Washington, powered by ArchiveSocial

3. Non-Print Legal Deposit: UK territoriality and social media sites

The British Library started web archiving in 2004, selectively archiving UK websites under licence. This has resulted in the Open UK Web Archive (http://www.webarchive.org.uk/), a collection to date containing approximately 15,000 websites. There is only a limited amount of social media content from Twitter, Facebook and YouTube in the UK Web Archive. It often form part of a "special collection", which is a group of websites brought together on a particular theme or an event, usually just archived for a fixed period of time. The decision not to systematically archive social media was related to the selective nature of the archive itself, and to resources constraints at the Library. The approach was to focus on representative websites according to our collection development policy (The British Library 2010). An example of these are Twitter pages belonging to the Prospective Parliamentary Candidates (PPCs) for the UK General Election 2010 (UK Web Archive, 2010). Even the limited exemplar content within the UK Web Archive however often required highly skilled technical staff to develop customised solutions outside our standard workflow. Hockx et. al

give a full account of the extensive effort and challenge involved in archiving Antony Gormely's public arts project One and Other including over 2000 hours of video streamed over the Real Time Messaging Protocol (RTMP), a non-HTTP protocol which the web crawler is unable to deal with. (Hockx-Yu, et al. 2010)

Legal Deposit has been a part of English law since 1662. It requires a copy of every UK print publication be given to the British Library by its publishers, and to five other major libraries that request it. Since 6 April 2013, Legal Deposit also covers material published digitally and online, including websites. Implementing Non-print Legal Deposit gave rise to the need for reconsidering the collection policy for social media, in relation to UK territoriality.

The Legal Deposit Regulations 2013 consider an online work as "published in the UK" and therefore in scope, if it meets either of the following criteria:

> (a) it is made available to the public from a website with a domain name which relates to the United Kingdom or to a place within the United Kingdom; or
>
> (b) it is made available to the public by a person and any of that person's activities relating to the creation or the publication of the work take place within the United Kingdom.
>
> (The Legal Deposit Libraries (Non-Print Works) Regulations, 2013)

The major social network platforms, YouTube, Twitter and Facebook, all use .com domain names so clearly do not meet the first criteria. YouTube is in fact out of scope as the Regulations do not apply to works consisting solely or predominantly of film or recorded sound (or both), where other forms of content are purely incidental. UK-published content in online sites such as LoveFilm, YouTube or Spotify are primarily in the form of sound and film, as defined by the Copyright Designs and Patents Act 1988[2], so is not covered by Non-print Legal Deposit.

Determining the territoriality of Twitter and Facebook however is less straightforward. As both contain content contributed by users around the globe, the second territoriality criteria cited above would only apply to UK-based individual or organisations, which does not warrant archiving twitter.com or facebook.com in their entirety. Once this is determined, it is still quite a challenge to identify technically and archive at scale just the content contributed by UK-based individuals and organisations. The Regulations do not define explicitly what constitutes "takes place within the United Kingdom", making it difficult to develop scalable technical solutions for discovering such content on the Web. A mixture of automated and manual means is currently used for this, including UK postal address, Geo-IP location and who-is records.

From the perspective of collection policy, social media are not treated differently from any other online work published in the UK. Regardless of the platform used for publication, non-print work is collected for Legal Deposit if it is fulfils the territoriality criteria. Our policy

[2] A sound recording is "a recording of sounds, from which the sounds may be reproduced, or a recording of the whole or any part of a literary, dramatic or musical work from which sounds reproducing the work or part may be produced, regardless of the medium on which the recording is made or the method by which the sounds are reproduced or produced" and a film is a "recording on any medium from which a moving image may by any means be produced". http://www.legislation.gov.uk/ukpga/1988/48/contents

also allows room for professional judgement, to deal with self-evident cases which fail all the manual or automated territoriality tests but are clearly created by an individual or organisation from the UK. An example of this is David Cameron's Twitter page, with a clear statement confirming that it is the official twitter site of the UK Prime Minister.

In-scope social media content will continue to be archived together with the rest of the UK web for non-print Legal Deposit, both as part of our broad annual domain crawls as well as additional collections reflecting national events or topics of common interest. Due to the difficulty in identifying automatically the UK content on non .uk domains, the coverage of social media content will remain limited until scalable solutions are developed for UK territoriality.

4. "Social archiving"?

The scale and complexity related to social media make it impossible for individual memory institutions to undertake alone the task of archiving and long-term preservation. The scope of what we are expected to collect will continue to grow and it is not feasible to collect and keep everything. While national libraries and archives should continue with developing solutions to collecting and safeguarding each nation's collective digital heritage, it is also important to realise that the geographical boundaries are somewhat arbitrary in the context of the web, where everything is intrinsically linked. When researchers study the live and the historical web, they would expect to find material relevant to research questions at hand, rather than where they come from or where the archived content was once hosted. This seems to introduce a dilemma: while resources are scarce to collect the increasing amount of digital content, there is a growing expectation from researchers for memory institutions to capture and provide access to the digital universe.

Maybe the distributed and highly participatory nature of social media itself might lend itself to developing new models for archiving. Crowd sourcing is not a new concept in the cultural heritage sector and its potential has already been evidenced by a number of successful projects. A recent inspiring initiative is the Boston Marathon Archive project (http://marathon.neu.edu/) hosted at the US Northwestern University, calling for contributions to a crowd-sourced archive of pictures, videos, stories and social media related to the Boston Marathon bombing on April 15, 2013. The Archive now has thousands of items, with items added each day, serving as a long-term memorial of a historical event. In the early stages the project went beyond requesting help from the crowd with data entry, but also asked for programming effort in building plugins for an open-source publishing platform (Omeka) to support the incorporation of various types of data.

In comparison with some at-risk resources on the Web, public social media content is currently archived reasonably well by a combination of researchers, companies and individuals using a wide range of commercial or open source tools. Twitter is perhaps the most remarkable in how they have put in place a range of services to ensure the longevity and availability of historical data. In addition to donating its digital archive of public tweets to the Library of Congress since 2010, Twitter rolled out a service allowing users to download their entire archive of tweets in late 2012 and introduced the Data Grants pilot project in 2014, through which access to public and historical data is given to selected research institutions (Twitter Blog, 2014). In addition they recently acquired Gnip, previously one of Twitter's data reseller partners, to further develop paid-for data products.

Archiving social media or digital heritage in general should not be the mission solely for memory institutions. Archivists, librarians, technologists, researchers, publishers, representatives of the government and industry have to work together to develop new models and tools appropriate to archiving and exploring social media. It is plausible to envisage a future collaborative model where a range of archiving services exists, allowing responsibilities and workload to be shared among stakeholders. As more and more scholars start using social media to study our world today, new models of collecting and donating data for research will emerge, overcoming some of the barriers currently preventing them from using such data.

Acknowledgments

I would like to thank Dr Andrew Jackson, Roger Coram and Dr Peter Webster, members of the British Library Web Archiving Team, for useful discussion and verification of technical and curatorial details which enabled this paper.

References

Archive-It, 2014. *Archiving Social Networking Sites with Archive-It.*
Available at: https://webarchive.jira.com/wiki/pages/viewpage.action?pageId=3113092
[Accessed 6 June 2014].

Archive-It, 2014. *Introduction to Umbra.*
Available at: https://webarchive.jira.com/wiki/display/ARIH/Introduction+to+Umbra
[Accessed 6 June 2014].

The British Library, 2010. *The British Library Collection Development Policy for Websites.*
Available at: http://www.bl.uk/reshelp/pdfs/modbritcdpwebsites.pdf
[Accessed 9 June 2014].

Boyd D; Ellison N, 2007. Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication,* 13(1), pp. 210-230. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/full
[Accessed 4 June 2014]

Hockx-Yu, H; Crawford, L; Coram, R; Johonson, S, 2010. *Capturing and replaying streaming media in a web archive - a British Library case study.*
Available at: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf
[Accessed 9 June 2014].

The Legal Deposit Libraries (Non-Print Works) Regulations.
Available at: http://www.legislation.gov.uk/uksi/2013/777/pdfs/uksi_20130777_en.pdf
[Accessed 9 June 2014].

Library of Congress, 2013. *Update on the Twitter Archive at the Library of Congress.*
Available at: http://http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf
[Accessed 5 June 2014].

The National Archives, 2014. *UK Government Web Archive captures official tweets and videos.*
Available at: http://www.nationalarchives.gov.uk/news/929.htm
[Accessed 6 June 2014].

Pennock, M., 2013. *Web-Archiving: DPC Technology Watch Report 13-01*, Digital Preservation Coalition. Available at http://dx.doi.org/10.7207/twr13-01
[Accessed 4 June 2014]

Pop R; Vasile G; Masanes J, 2010. *Archiving Web Video.* Vienna, Proceedings of the International Web Archiving Workshop IWAW 2010, 00.42-28.
Available at: http://iwaw.europarchive.org/10/IWAW2010.pdf
[Accessed 5 June 2014].

Snohomish County, Washington - Social Media Archive.
Available at: http://snoco.wa.gov.archivesocial.com/
[Accessed 9 June 2014].

Technical Committee ISO/TC 46, 2012. *Information and documentation — Statistics and Quality Indicators for Web Archiving.*
Available at: http://netpreserve.org/sites/default/files/resources/SO_TR_14873__E__2012-10-02_DRAFT.pdf
[Accessed 5 June 2014].

Twitter Blog, 2014. *Introducing Twitter Data Grants.*
Available at: https://blog.twitter.com/2014/introducing-twitter-data-grants
[Accessed 9 June 2014].

UK Web Archive, 2010. *UK General Election 2010: Candidates.*
Available at: http://www.webarchive.org.uk/ukwa/collection/35389445/page/1
[Accessed 9 June 2014].