Preserving Social Media: applying principles of digital preservation to social media archiving

Sara Day Thomson
Research Officer, Digital Preservation Coalition
sara.thomson@dpconline.org
dpconline.org

Abstract

While new research in social media holds great potential for new understanding, in the rush to exploit this new source of data we must not overlook the importance of ensuring long-term access. The issues of archiving social media—an extension of web archiving—reflect many of the challenges addressed by digital preservation, including external dependencies, personal data and individual privacy, ownership, and scalability. This is not surprising considering the capture of social media is essentially an act of digital preservation – taking action to ensure future access to otherwise temporary or vulnerable digital material. This paper examines a few key challenges to archiving social media, positioning them within the larger framework of challenges facing digital preservation. Each section discusses how established practices in digital preservation can be directly applied to social media data. This paper does not provide a detailed examination of individual platforms, or their data sharing policies, but rather provides an overview of the challenges to preserving social media and how those challenges might be met by approaches developed by the digital preservation community. This paper invites scholars and researchers, in academe and industry, to think about their data from the long view. Simultaneously, it encourages information practitioners (research data managers, librarians, archivists) to consider the needs of researchers (and future researchers). It lays out a path to collaboration to build a more sustainable approach to the preservation of social media.

Introduction: Lessons from the (short) History of Digital Preservation

In a report released in 2013, the UK Data Forum projected growing importance for social media research and released a call for further development:

> 'Through social media, millions of human interactions occur and are recorded daily, creating massive data resources which have the potential to aid our understanding of patterns of social behaviour … . Social media analytics represent an opportunity to invest in extensive, large scale social research that is within a temporal frame that cannot be achieved through either snapshot surveys or interviews or via longitudinal panel research' (UK Data Forum, 2013).

Through this statement, the UK Data Forum characterises a trend in social media research in social science and economics but also computer science, marketing, and a quickly broadening range of other disciplines (Weller and Kinder-Kurlanda, 2016). Researchers and research institutes across the UK (and the rest of the

world) eagerly pursue new insights revealed by this new and novel form of data. This burst of activity in social media research mirrors analogous trends in the corporate sector. Commercial institutions increasingly look to social media data to collect and analyse information about demographic groups and individual consumers (ADMA, 2013).

While new research in social media holds great potential for new understanding, in the rush to exploit this new source of data we must not overlook the importance of ensuring long-term access. The issues of archiving social media—an extension of web archiving—reflect many of the challenges addressed by digital preservation, including external dependencies, personal data and individual privacy, ownership, and scalability (DPC, 2016e). This is not surprising considering the capture of social media is essentially an act of digital preservation – taking action to ensure future access to otherwise temporary or vulnerable digital material. This shared enterprise has existed between web archiving and digital preservation for a long time. This relationship was recently publicly lauded with the award of the first ever Digital Preservation Coalition Fellowship for lifetime achievement in digital preservation to Brewster Kahle of the Internet Archive (DPC, 2016a).

However, despite this long relationship, the future of research using social media data depends on current professionals addressing the issues of long-term preservation. While the more immediate difficulties of access and use often dominate the conversation about social media data, the issues of long-term preservation also demand immediate attention. Most forms of social media content are vulnerable to loss and owned by platforms with no contractual or legal requirement to preserve individual users' data. Digital preservation enables reproducible research; without preserving social media, research cannot be reproduced.

This paper examines a few key challenges to archiving social media, positioning them within the larger framework of challenges facing digital preservation. While drawing on a number of resources from the digital preservation community, this paper presents concepts and accepted practice as described in the *Digital Preservation Handbook* available on the Digital Preservation Coalition website (http://dpconline.org/handbook). Each section discusses how established practices in digital preservation can be directly applied to social media data. From defining a 'Designated Community' to 'bit preservation', hard-earned best practice in digital preservation is just as relevant to social media as to eBooks or PDFs. It borrows lessons from other sectors, such as the parent practice of web archiving to the emerging field of video game preservation, in order to suggest creative solutions. This paper does not provide a detailed examination of individual platforms, or their data sharing policies, but rather provides an overview of the challenges to preserving social media and how those challenges might be met by approaches developed by the digital preservation community. This paper invites scholars and researchers, in academe and industry, to think about their data from the long view. Simultaneously, it encourages information practitioners (research data managers, librarians, archivists) to consider the needs of researchers (and future

researchers). It lays out a path to collaboration to build a more sustainable approach to the preservation of social media.

Who Is the Future?: Defining a Designated Community for Social Media Data

The term 'Designated Community' comes from the Open Archival Information System (OAIS) Reference Model (or ISO 14721:2012) and refers to a repository's primary users. OAIS has played an important role in the development of digital preservation strategies but has also been criticised by practitioners for failure to support the practicalities of implementation in certain situations, such as web archives (Rosenthal, 2015). Even the concept of Designated Community as it is presented in the OAIS model has been criticised for a lack of inclusivity (Bettivia, 2016). Despite these identified shortcomings of OAIS, the concept of the Designated Community maintains a central role in the planning and development of many successful digital preservation programmes. Best practice encourages organisations to consider the 'beginning' of digital preservation to be the point of creation so that necessary understanding about how the object will be used and by whom can be captured (DPC, 2016b). Articulating the Designated Community, and its requirements, provides the cornerstone on which many other components of a digital preservation strategy depend. For example, a Designated Community holds specialist knowledge about a subject, making them a good source of information about contextual information needed to make the data intelligible. Thus, defining and monitoring the Designated Community helps to define metadata requirements. Similarly, developing an approach to preserving social media will be strengthened by an understanding of who will use the preserved data and how. New and novel forms of data, such as social media, have an even greater need for a clearly defined Designated Community than more established forms of data. In the absence of best practice and professional standards, user requirements will help determine the formats, metadata, and modes of access to ensure the preserved content will remain useful and reliable in the future.

The ways a social media collection will be used will help collecting institution to shape how the data is collected, in what form, and with what types of metadata. For example, the UK Data Service and the UK Web Archive both capture Twitter data, but the collections and their metadata look very different. The UK Data Service defines its main user community as 'social science data users within higher education (HE) and further education (FE) in the UK' and that the Archive will provide 'formats suitable for research, teaching or learning' (UKDA, 2016). An example of a Twitter dataset held by the UKDS is the 'After Woolwich Twitter corpus', comprising the tweets harvested as part of a project that analysed social reactions to the murder of Drummer Lee Rigby in Woolwich in 2013 (Innes, 2016). The available data includes tweet IDs and the Python code used to download the tweet data from Twitter. The tweet IDs allow other researchers to request the same data directly from the Twitter API, which will provide them with machine-readable JSON files required for computational analytics. The data from the API supplies rich metadata, which will maximize the types of queries and analysis that can be performed. While sharing tweet IDs does not fulfil

professional research data standards, Twitter's Developer Agreement and Policy currently forbids the sharing of Twitter research data in a form that meets these standards (Twitter, 2016; Weller and Kinder-Kurlanda, 2015). By contrast, the UK Web Archive states that their collections are aimed at users 'across a wide spectrum of interest and knowledge: the general reader, the teacher, the journalist, the policy maker, the academic and personal researcher, and many more besides' (UK Web Archive). This user community encompasses a much broader range of people and does not target specialists or those with the computational skills to perform analytics on a dataset. The archived Twitter pages made through the UK Web Archive are harvested by a web crawler and displayed in a similar format to live Twitter (using the Wayback Machine). This example of the archived Twitter feed of Nick Clegg taken on 25 January 2014 shows a snapshot of the former Lib Dems leader's activity on the social media platform: https://www.webarchive.org.uk/wayback/archive/20140125085458/https://twitter.com/nick_clegg . These harvested pages are not designed for computational analytics but rather to enable future users to browse Twitter activity from a particular point in time.

The long-term preservation of social media data faces a considerable number of challenges, mostly from the restrictions imposed by platforms and the ethical dilemmas of re-using individuals' social interactions without their awareness (Thomson, 2016; Voss, et. al., 2017). These challenges demand that researchers and archivists must make difficult compromises when it comes to capturing and preserving these valuable records. Researchers and archivists must work together to closely define the Designated Communities who will most likely use social media collections to ensure that future users will have access to something meaningful and usable.

(Re-)Selection, Risk Assessment, and Policy Developing for Futureproof Social Media

Making compromises about what to preserve for the long-term is not unique to social media data. For example, many forms of digital content require large amounts of storage space that collecting institutions simply do not have. Digital images and video quickly fill up precious server space and continue to grow at a steady rate (Wright, 2012). In addition to practical concerns, the possibility of personal data or even sensitive data within digital content often requires redaction or disposal. Digital preservation specialists have been developing strategies and best practice for decades to aid in the process of making decisions about what to save, what to discard, and how to economize storage (Cornell, 2016). Accepting new content relies on a selection process based on collection policies, resources, risk assessment, and institutional buy-in (DPC, 2016e). Long-term preservation relies on these same factors, but in the context of a much longer timescale. For example, while in the short-term issues such as storage media and file format may not have a significant impact, in the medium to long term, these issues are critical because of media obsolescence, decay, and failure (DPC, 2016e). It may be helpful to think of this secondary process as 're-selection', although making decisions about what content to save for the long-term should ideally be made when new content is accepted for deposit.

Re-selection for social media data could benefit from two digital preservation approaches in particular: risk assessment and policy development. Evaluating social media collections through systematic risk assessment will support the development of a strategy for long-term preservation. Logging these risks in a register where they can be reviewed by other staff, including management, will help to shape institutional policies. A 2011 study showed that institutions that have a mandate for digital preservation established in institutional policy have more successful digital preservation programmes (Sinclair, et. al.). Risk assessment and policy development work together to elevate digital preservation from ad hoc project work to routine, efficient, and sustainable business-as-usual. Risk assessment and policy development can also help standardize (and democratize) the capture and preservation of social media data, a practice which is currently bespoke, ad hoc, and only available to a specialist and economic elite (Weller and Kinder-Kurlanda, 2015; Puschmann, et. al., 2014).

There is not space in this paper to present a full assessment of the risks to the long-term preservation of social media data. However, two generic risks associated with these data demonstrate how such an assessment might help shape institutional policy and improve the preservation of social media. First: the ethics of re-using data created by users who may be unaware of how their data is being used, and second: the competing risk of *not* acting to preserve social media data (Thomson, 2015).

The risk of re-using social media data for research or heritage collections when social media users may be unaware of how their data is being used could lead to several different types of consequences. Without implementing *any* mitigating steps, institutions could risk:

- accidental disclosure of personal information or even sensitive information causing potential harm to individuals
- damage to relations with users and with the public
- damage to institutional reputation and possible loss of funding
- lawsuit and resulting fines for breach of data protection regulations

The dangers posed by re-using social media data, however, are not unsurmountable. By implementing mitigating steps, institutions can reduce the likelihood and impact of the consequences listed above. Guidance for the ethical treatment of personal can be found through the OECD (http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm), the ESRC (http://www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics/), and the Association of Internet Researchers (AoIR) (http://aoir.org/ethics). The AoIR, for example, provides a list of questions to pose about data to identify the ways individuals might be revealed or made vulnerable (Markham et. al., 2012). By removing identifying information from social media data, such as legal names or place names, an institution demonstrates a commitment to protect the individuals represented in the data. Restricting access to social media data collections to only registered or credentialed users also reduces risk.

Institutions might also consider publishing a statement about how it uses social media data and the benefits this data provides not only for researchers and collecting institutions but for the public. Transparency about data practices and communication about the benefits of preserving social media also mitigate the risk of social media users being unaware of how their data is used.

While the risk of re-using social media data is caused by actively capturing and preserving social media, *failing* to preserve social media also constitutes a risk with potentially irreversible consequences. Social media platforms – such as Facebook, Twitter, and LinkedIn – do not have any contractual obligation to preserve user data for the long-term. Most data policies provide information on how users can delete content or request that personal data held by the platform be deleted, but no platforms agree to act as a long-term repository for user data[1]. For instance, the Twitter Terms of Service states:
'The Twitter Entities shall not be liable for … any data loss'. The LinkedIn User Agreement states: 'LinkedIn is not a storage service. You agree that we have no obligation to store, maintain or provide you a copy of any content or information that you or others provide'. If researchers and collecting institutions do not capture and preserve social media data, it is likely these data will disappear and this valuable cultural resource will be lost. Institutions with a remit to preserve research data, government data, or cultural heritage data can mitigate the risk of this loss through capturing and preserving social media data as a routine, adequately-funded, and organisationally-supported activity. Through articulating these risks and implementing a responsive strategy, institutions can prevent the catastrophic loss of user data. These actions will also support advocacy and campaigning for more open social media policies for use by research and heritage institutions.

Future-telling: Bit Preservation, Metadata, and Documentation

Bit preservation refers to the storage and maintenance of the bit streams that underlie all digital content – the 1s and 0s that encode the information contained within files and groups of files (DPC, 2016c). While some practitioners might argue that digital preservation has moved beyond bit preservation to more complex issues, maintaining access to digital code remains an unwavering priority when it comes to ensuring long-term access (Rosenthal, 2010). The concept of bit preservation and its centrality to maintaining long-term access illustrates one of the fundamental characteristics of preservation. Digital preservation prioritises maintaining the digital object over the pressures of immediate access. A preservationist attitude takes the view that if an object is not preserved, it cannot be made accessible. The opposing attitude would argue that if something is not being accessed, there is no point in preserving it. However, in some cases, a digital object may not be available for immediate access because of legal or practical reasons (as opposed to technological ones) but possesses great historical or informational value.

[1] Facebook Data Policy (29 September 2016): https://en-gb.facebook.com/policy.php; Twitter Terms of Service (30 September 2016): https://twitter.com/tos?lang=en; LinkedIn User Agreement (23 October 2014): https://www.linkedin.com/legal/user-agreement

An illustration of this situation from the world of paper records is the '50 Year Rule' introduced with the Public Records Act in 1958, a policy that required UK government to transfer records to the National Archives to release to the public after 50 years (TNA, 2009). In 1967, this release period was reduced to 30 years and since the increase in digital records and the implementation of FOI, authorities have now reduced this period even further (TNA, 2015). This means that before 1967, government offices expended the effort and resources to preserve records for up to 50 years despite the fact that users outside of that agency could not access them. As is the case with UK government records, the preservation of the underlying bits of social media data is crucial to ensuring future users can access and use these records one day when conditions change.

Metadata and documentation are also critical to interpreting the underlying bits of a digital object itself. In fact, oftentimes, metadata and documentation associated with an object are treated as digital objects in their own right. Standards like METS – an encoding standard to encapsulate a digital object and its metadata – demonstrate that institutions view information *about* an object (such creation date, software dependencies, rights issues, etc.) are just as important as the digital object itself (Cantara, 2005). The digital preservation strategy of emulation presents a particularly compelling case for the preservation of metadata and documentation (DPC, 2016d). An emulation strategy aims to preserve digital content by maintaining documentation about the technology (hardware and software) required to interpret the digital object rather than migrating (or updating) the object to a newer format that can be opened and read with newer versions of technology (DPC, 2016c). Emulation preserves a digital object in its most authentic form, relying on the ability to re-create ('emulate') the technology needed to open (and interpret) that object in the future. Emulation is a popular strategy, for example, for the preservation of computer and video games (Conley, et. al., 2004). Emulation allows future players to experience the game more closely to how it was experienced by contemporary players (Conley, et. al., 2004). For a game to be emulated successfully, however, information about the original software or console used to play the game must be available and understandable (Kraus, et. al, 2012). Metadata and documentation, in this case, are equally as important as the object itself.

Much like a video game, social media is an experience for users. The content itself, for example the text of a tweet, does not represent the full meaning of a social media exchange on a platform (Dijck, 2013). The number of likes and who has liked it, for instance, directly affects how influential a tweet is within a network (Dijck, 2013). The ability to reply or re-tweet also contributes significantly to the experience of the participatory, interactive medium of Twitter (Dijck, 2013). Some uses of social media require the original conditions of the platform to an even greater degree to convey meaning. The Instagram artist Amalia Ulman, for example, uses colour schemes and other interface characteristics of the desktop browser version of the platform to create works of art (http://webenact.rhizome.org/excellences-and-perfections). Captured using Rhizome's Webrecorder.io tool and hosted on Webenact, a server for storing captures from the live web, the Instagram art created by Ulman can be displayed using the same version of Instagram

used to create the work (https://webrecorder.io/_faq). Trying to view the live account using updated interface designs would destroy the proportions and colour choices made by the artist. Webrecorder.io prevents the loss of artworks like Ulman's by capturing the object and its environment, allowing future users to experience the artists' interactive work as closely as possible to the original. Of course, Webrecorder.io cannot control the devices used by future users to access the web-based art. However, by capturing and preserving the social media artworks and their contextual information in archive-friendly WARC files (Littman, et. al., 2016), future users will have the ability to 're-play' the artworks in as an authentic form as possible. As illustrated in these examples, the digital preservation approach of emulation could provide a meaningful and authentic mode of preservation for social media. Emulation will only be effective if the 'bits' of social media data – as well as the metadata and documentation required to understand it in its 'original habitat' – are captured and preserved.

The Chicken or the Egg: The Vital Role of Collaboration and Advocacy for the Survival of Social Media Research and Heritage Collections

As first addressed in the discussion about Designated Communities, the relationship between researchers and information professionals will have a significant impact on the survival of social media for long-term access. To ensure long-term access to these data and its contexts, information professionals need to understand how it is used and how it is likely to be used in the future. The best way for them to acquire this understanding is for researchers to produce an abundance of diverse research using social media. Thus, the preservation of social media records faces the first of several 'chicken or the egg' conundrums. Which comes first: the growth of research using social media or better standards for social media data? Although the current lack of best practice for social media research data poses a challenge for researchers (Weller and Kinder-Kurlanda, 2015), information professionals and the administrators and managers who support research and collecting institutions can aid the growth of social media research in other ways. Information professionals, administrators, and institution leaders can aid growth by advocating for wider access to archived social media and by proactively engaging with users to advise on issues of data curation and management as they arise.

The path forward for information professionals and institutional leaders faces another chicken or the egg conundrum. Which comes first: campaign for more open platform policies or promote the publication of social media research that demonstrates the importance of this type of research? While platform restrictions and the complexity of managing social media data hamper efforts to encourage growth, information professionals and their institutions must be resilient and think creatively. On the one hand, institutions must 'be the chicken' and make a case to policy-makers for the importance of archiving this data, a feat which will require the demonstration of compelling research well-supported by data management. On the other hand, researchers and information professionals must 'be the egg' and show

how this research has been possible *despite* restrictive platform policies. They must show how much more robust and democratic this research could be if better policies could be negotiated with platforms.

The road to a data access framework more conducive to reproducible research starts with a strong working relationship between researchers and information professionals. Close collaboration between researchers and data managers and archivists creates a solid foundation for advocacy within institutions that have a stake in social media research and heritage collections. The practice of digital preservation has relied significantly on internal advocacy and as a result has developed a wealth of guidance on building a business case and for developing stakeholder buy-in. The Digital Preservation Handbook devotes an entire section to this topic: 'Business cases, benefits, costs, and impact' (http://dpconline.org/handbook/institutional-strategies/business-cases-benefits-costs-and-impact). These strategies also provide useful guidance for social media research and heritage collections. By performing systematic risk assessments and recording those assessments alongside relevant benefits, researchers and information professionals will have a valuable tool at their disposal. These resources – a risk register, documentation of benefits, and persuasive business case – will encourage the integration of social media research objectives and collection strategies into institutional policy. The creation of policies for capturing and preserving social media data will lead to better-supported social media research and to richer heritage collections for future generations.

As the institutional practices for social media research and collecting become more prevalent, research and collecting organisations will have a stronger case for external advocacy. The Digital Preservation Handbook provides guidance on how to shape and encourage advocacy internally and also to a wider audience. Two overriding approaches in particular are also well-suited to support advocacy for social media. One, by communicating the methods and benefits of social media research and heritage collections, researchers and information professionals can build trust and support among the public. Two, by aligning the preservation of social media with national narratives and political objectives, social media researchers and information professionals can assert the importance of ensuring long-term access to these valuable data. Both strategies benefit from both small and grand actions: from individual research projects to the institutional policies of national organisations.

Communicating the benefits of social media research, and simultaneously providing reassurance about the safe treatment of personal data, will go a long way in garnering public support. In 2014, Ipsos MORI, supported by the ONS and ESRC, published a report on public attitudes towards using linked administrative data in research based on dialogues with members of the public (Cameron, et. al.,2014). While the report did not address attitudes towards social media data, linked administrative data is an analogous form of data – created through routine business and *not* generated deliberately for the purposes of research. Both linked administrative data (such as health records and student records) and social media data can provide a rich source of 'native' data for researchers to make observations and discover patterns of behaviour. The Ipsos MORI study found that: 'At the beginning of the dialogues, low awareness of the uses of social

research drove scepticism about its value' (Cameron, et. al., 2014). However, the report also stated: 'Later in the dialogues, when participants had learned more about the aims and methods of social research, they tended to be more positive about its value' (Cameron, et. al., 2014). Researchers and information professionals could also engender greater public support through a coordinated communications campaign that addresses social media data practices and the benefits of social media research and heritage collections.

By aligning these messages about social media research and heritage collections with national narratives and political objectives, social media campaigners will likely achieve better traction. The DocNow project provides just such an example of social media researchers and information professionals coming together to assert the importance of social media preservation on a national scale (http://www.docnow.io/). Inspired by the activity on Twitter following the police shooting of Michael Brown in Ferguson, Missouri in 2014, archivists and researchers developed the DocNow ('Document the Now') project to capture these responses for their historical and social value (Wortham, 2016). With an overt focus on ethical data practices, the DocNow team aims to enable researchers, scholars, and archivists to capture and preserve social media responses to major national events (http://www.docnow.io/). This approach makes use of one of social media's most compelling characteristics: a real-time barometer of online social responses to events across the globe. Responses on social media are something a large percentage of the population can relate to. A Pew Research Centre study in 2016 estimated that 69% of Americans use some type of social media (2017). A wearesocial 2017 special report published that around 2.8 billion people across the globe use social media, or about 37%. The importance of social media to the lives of many people around the world provides a compelling basis for the argument to ensure long-term preservation of social media as part of research and heritage collections. The research and information practitioner community would benefit from leveraging this argument to campaign for more open platform policies for accessing and sharing social media data.

The Best Way Forward: Conclusions and Recommendations

As progress in digital preservation has shown, convincing non-specialists of the importance of protecting our digital legacy is not a sprint but a marathon. While advocacy and public relations can take a long time (and repeated reinforcement), social media data requires immediate action to ensure long-term access. Like many digital preservation strategies, the best way forward is a twofold path. On the one hand, practitioners must make compromises and maximize the limited options available to prevent loss of social media data. On the other hand, practitioners and other stakeholders will have to advocate and campaign first, for increased support from institutional leaders and policy-makers, and second, for more open platform data policies. Platform restrictions and a lack of understanding about data practices currently stymy the preservation of social media. However, by persevering with creative solutions and building open communities of exchange and mutual support, researchers and information professionals can make

important progress in safeguarding an important cultural legacy for future researchers, policy-makers, and good citizens.

Works Cited

Association for Data-driven Marketing and Advertising (ADMA), 2013, 'Best Practice Guideline: Big Data', http://www.admaknowledgelab.com.au/compliance/compliance-help/general/data-and-privacy/codesand-guides/best-practice-guideline-big-data

Bettivia, R, 2016, 'The power of imaginary users: Designated communities in the OAIS reference model', Proceedings of the Association for Information Science and Technology, DOI: 10.1002/pra2.2016.14505301038

Cameron, D, Pope, S, and Clemence, M, 2014, 'Dialogue on Data: Exploring the public's views on using administrative data for research purposes', Ipsos MORI Social Research Institute, http://www.esrc.ac.uk/files/public-engagement/public-dialogues/dialogue-on-data-exploring-the-public-s-views-on-using-linked-administrative-data-for-research-purposes/

Cantara, L, 2005, 'METS: The Metadata Encoding and Transmission Standard', *Cataloging & Classification Quarterly*, Vol. 40 Iss. 3-4, pp. 237-253, DOI: 10.1300/J104v40n03_11

Conley, J, Andros, E, Priti, C, Lipkowitz, E, and Perez, D, 2004, 'Use of a Game Over: Emulation and the Video Game Industry', *Northwestern Journal of Technology and Intellectual Property*, Vol. 2 Iss. 2, http://scholarlycommons.law.northwestern.edu/njtip/vol2/iss2/3/

Cornell University, last updated 2 March 2016, 'Digital Content Review: Process and Results', *Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions*, online tutorial developed and maintained by Cornell University Library, 2003-2006; ICPSR, 2007-2012; MIT Libraries, 2012-present, http://dpworkshop.org/workshops/management-tools/process-results

Digital Preservation Coalition (DPC), 2016a, 'Digital Preservation Awards 2016 – Winners Announced!', http://dpconline.org/advocacy/awards/2016-digital-preservation-awards

Digital Preservation Coalition (DPC), 2016b, 'Creating digital materials', *Digital Preservation Handbook*, http://dpconline.org/handbook/organisational-activities/creating-digital-materials

Digital Preservation Coalition (DPC), 2016c, 'Glossary', *Digital Preservation Handbook*, http://dpconline.org/handbook/glossary

Digital Preservation Coalition (DPC), 2016d, 'Preservation action', *Digital Preservation Handbook*, http://dpconline.org/handbook/organisational-activities/preservation-action#ref

Digital Preservation Coalition (DPC), 2016e, 'Preservation Issues', *Digital Preservation Handbook*, http://handbook.dpconline.org/digital-preservation/preservation-issues

Dijck, J van, 2013, The Culture of Connectivity: A Critical History of Social Media, Oxford Scholarship Online, e-Book, DOI:10.1093/acprof:oso/9780199970773.001.0001

Innes, M (Cardiff University), 2016, Dataset: 'After Woolwich Twitter corpus', PI: 10.5255/UKDA-SN-852078, https://discover.ukdataservice.ac.uk/catalogue/?sn=852078&type=Data%20catalogue

Kraus, K, and Donahue, R, 2012, '"Do You Want to Save Your Progress?": The Role of Professional and Player Communities in Preserving Virtual Worlds', *Digital Humanities Quarterly*, Vol. 6 Iss. 2, http://www.digitalhumanities.org/dhq/vol/6/2/000129/000129.html#d13192e374

Littman, J, Chudnov, D, Kerchner, D, Peterson, C, Tan, Y, Trent, R, Vij, R, and Wrubel, L, 2016, 'API-based social media collecting as a form of web archiving', International Journal of Digital Libraries, First Online 28 December 2016, DOI: 10.1007/s00799-016-0201-7

Markham, A, and Buchanan, E, 2012, 'Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee' (Version 2.0), Association of Internet Researchers (AoIR), https://aoir.org/reports/ethics2.pdf

National Archives (TNA), 2009, '30 Year Rule Review', http://webarchive.nationalarchives.gov.uk/20090516124148/http:/www.30yearrulereview.org.uk/background.htm

National Archives (TNA), 2015 , '20-year rule', http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-projects/20-year-rule/

Pew Research Centre, 12 January 2017, 'Social Media Fact Sheet', http://www.pewinternet.org/fact-sheet/social-media/

Puschmann, C and Burgess, J, 2014, 'The Politics of Twitter Data', In K Weller *et al.* (Eds) *Twitter and Society*, New York: Peter Lang Publishing.

Rosenthal, D, 2010, 'Bit Preservation: A Solved Problem?', *International Journal of Digital Curation*, Iss. 1 Vol. 5, pp. 134-148, file:///C:/Users/sdaythomson/Downloads/148-689-1-PB.pdf

Rosenthal, D, October 2015, 'The case for a revision of OAIS, OAIS Community, DPC Wiki, http://wiki.dpconline.org/index.php?title=The_case_for_a_revision_of_OAIS
Thomson, S, 2016, *Preserving Social Media*, DPC Technology Watch Report, DOI: http://dx.doi.org/10.7207/twr16-01

Sinclair, P, Duckwork, J, Jardine, L, Keen, A, Sharpe, R, Billenness, C, Farquhar, A, and Humphreys, J, 2011, 'Are you Ready? Assessing Whether Organisations are Prepared for Digital Preservation', *International Journal of Digital Curation*, Iss. 1 Vol. 6, pp. 268-281, http://www.ijdc.net/index.php/ijdc/article/viewFile/178/247

Thomson, S, 2015, 'Preserving Social Media', a Digital Preservation Coalition *Technology Watch Report*, DOI: http://dx.doi.org/10.7207/twr16-01

Twitter, last updated 30 September 2016, 'Developer Agreement & Policy', https://dev.twitter.com/overview/terms/agreement-and-policy

UK Data Archive (UKDA), 15 June 2016, 'Preservation Policy', http://www.data-archive.ac.uk/media/514523/cd062-preservationpolicy.pdf

UK Data Forum, 2013, 'UK Strategy for Data Resources for Social and Economic Research', http://www.esrc.ac.uk/files/news-events-and-publications/news/2013/uk-strategy-for-data-resources-forsocial-and-economic-research/

UK Web Archive, 'Who Is the UK Web Archive For?', https://www.webarchive.org.uk/ukwa/info/about, last accessed 25 May 2017

Voss, A, Lvov, I, and Thomson, SD, 2017, 'Data Storage, Curation and Preservation', in *The SAGE Handbook of Social Media Research Methods*, Luke Sloan and Anabel Quan-Haase, Eds., Sage Publications Ltd., eISBN-13: 9781473987210

Wearesocial, Digital in 2017: Global Overview, https://wearesocial.com/special-reports/digital-in-2017-global-overview

Weller, K and Kinder-Kurlanda, K, 2015, 'Uncovering the Challenges in Collection, Sharing and Documentation: the Hidden Data of Social Media Research?', *Standards and Practices in Large-Scale Social Media Research: Papers from the 2015 ICWSM Workshop* http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/viewFile/10657/10552

Weller, K and Kinder-Kurlanda, K, 2016, 'A Manifesto for Data Sharing in Social Media Research', DOI: 10.1145/2908131.2908172, http://dl.acm.org/citation.cfm?doid=2908131.2908172

Wortham, J, 21 June 2016, 'How an Archive of the Internet Could Change History', The New York Times Magazine, https://www.nytimes.com/2016/06/26/magazine/how-an-archive-of-the-internet-could-change-history.html?_r=0

Wright, R, 2012, 'Preserving Moving Pictures and Sound', a Digital Preservation Coalition *Technology Watch Report*, DOI: http://dx.doi.org/10.7207/twr12-01