Preserving Social Media

Sara Day Thomson

DPC Technology Watch Report 16-01 February 2016

Series editors on behalf of the DPC
Charles Beagrie Ltd.

Principal Investigator for the Series
Neil Beagrie

DigitalPreservationCoalition

This report was supported by the Economic and Social Research Council [grant number ES/JO23477/1]

ISSN: 2048-7916

DOI: http://dx.doi.org/10.7207/twr16-01

First published in Great Britain in 2016.

Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow.

The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow. They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Sara Day Thomson. The report is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for, and managing editor of, this Series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on text prior to release: William Kilbride (Chair), Janet Delve (University of Portsmouth), Marc Fresko (Inforesight), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), and Dave Thompson (Wellcome Library).

Acknowledgements

I would like to thank the many researchers and practitioners who graciously contributed to this report through interviews, recommendations, and support. In particular, thank you to the UK Data Archive (UKDA) and UK Data Service (UKDS) for their initiative in commissioning this study and the ongoing guidance and encouragement from Nathan Cunningham, UKDA Associate Director and UKDS Functional Director for Big Data Network Support. I would also like to extend my appreciation to all of the practitioners – archivists, librarians, and data scientists – who contributed to the Focus Group that formed the early foundations of this report, and to Clémence Agostini and Dorota Walker from the British Library for their hospitality. Thank you in particular to Libby Bishop at UKDA for her input and recommendations on the ethics of big data research as well as to Garth Stewart from National Records of Scotland for his enthusiastic contributions. Thank you also to Ed Pinsent and Stephanie Taylor at ULCC for their early support and feedback and to Peter Webster for sharing his broad learning and insight.

A special thank you and acknowledgement to the practitioners who provided the interviews that stand as the backbone of this report, to Matthew Williams, Luke Sloan, and Pete Burnap at the Social Data Science Lab in Cardiff – thank you for your hospitality and ongoing feedback, to Clare Lanigan from Digital Repository Ireland for her keen expertise and generous support, to Katrin Weller from the GESIS Leibniz Institute for the Social Sciences in Cologne for sharing her wealth of experience and insight, and to Tom Storrar and Suzy Espley from The National Archives for sharing insight and wisdom from their valuable experiences developing the UK Government social media archives.

With a freshly earned appreciation, I would like to thank the innovative and resourceful scholars, teachers, practitioners, and developers who have tackled the giant task of preserving social media for the benefit of their communities and future users. Lastly, I would like to thank the DPC and my family for their support, patience, and friendship.

Sara Day Thomson
January 2016

Contents

1. Abstract 1
2. Executive Summary 2
3. Introduction 4
3.1. What is social media? 4
3.2. Who archives social media? 5
3.3. How is social media archived? 7
3.4. What are the challenges to preserving social media? 8
4. Strategies 9
4.1. Application Programming Interfaces (APIs) 9
4.2. Data resellers 11
4.3. Third-party services 12
4.4. Platform self-archiving services 12
4.5. Research dataset identifiers 13
4.6. Research documentation and metadata 13
5. Challenges 15
5.1. Platform terms and conditions 15
5.2. Copyright infringement 17
5.3. Brokering agreements with platforms 17
5.3.1. Twitter Research Access at the Library of Congress 18
5.3.2. MIT's Laboratory for Social Machines 19
5.4. Ethics of user privacy and awareness 20
5.5. Selection 22
5.6. Storage and search 24
5.7. Long-term preservation 25
6. Case Studies 28
6.1. Social Data Science Lab and the COSMOS Platform, Cardiff University 28
6.2. Social Repository of Ireland feasibility study (Insight @ NUI Galway & The Digital Repository of Ireland) 29
6.3. GESIS Leibniz Institute for the Social Sciences 30
6.4. The National Archives' UK Government Social Media Archive 31
7. Conclusions and Recommendations 33
8. Glossary 35
9. Further Reading 37
9.1. Case studies 37
9.2. Tools and resources 37
10. References 39

1. Abstract

Social media plays an increasingly important role as we embrace networked platforms and applications in our everyday lives. The interactions of users on these web-based platforms leave valuable traces of human communication and behaviour revealed by ever more sophisticated computational analytics. This trace – the data generated by social media users – is a valuable resource for researchers and an important cultural record of life in the 21st century. As the programming and infrastructures of social media, or Web 2.0, mature and grow, researchers and collecting institutions need new techniques for capturing this web-based content. This report provides an overview of strategies for the archiving of social media for long-term access, for both policy and implementation. Specifically, it addresses social networking platforms and platforms with significant amounts of user-generated content, excluding blogs, trading, and marketing sites, which are covered in other *Technology Watch Reports*. Alongside the potential strategies for archiving social media, the challenges facing its preservation by non-commercial institutions will be explored. In the absence of established standards and best practice, this report draws on recent initiatives undertaken by research, heritage, and government archives in the UK, Ireland, and Germany. Based on the current conditions surrounding social media data and the lessons demonstrated by a range of case studies, recommendations for the future development of social media preservation for research and heritage collections will be made.

This report is intended for any institution with an interest in preserving social media for supporting research, public records, or cultural heritage. Good practice in managing and archiving social media data applies to researchers, archivists, and librarians alike. As social media continues to grow as a source of official government and corporate communications, the importance of effective preservation will increase. Similarly, as more people across the globe replace traditional forms of communication and expression with the media available through social media platforms, the record of histories and cultures will increasingly rely on the ability of researchers and archivists to document this fast-paced and dynamic form of data.

2. Executive Summary

Web-Archiving by Maureen Pennock, a *Technology Watch Report* published in 2013, presented the issues and solutions for archiving the World Wide Web. The present report looks at the related issues of preserving social media, a form of Web 2.0 that poses particular challenges to web crawlers and traditional methods for archiving the Web. Furthermore, social media platforms provide large quantities of machine-readable data that support computational analytics, an emerging practice used by a number of academic disciplines, journalists, and other professionals. In order to harvest and preserve this type of rich data in a meaningful way, research and collecting institutions need new approaches and methods. These approaches will need to capture data and its affiliated context on a large scale, rather than copies or snapshots of web pages.

This report will map the landscape of preserving social media for long-term access by presenting practical solutions for harvesting and managing the data generated by the interactions of users on web-based networking platforms such as Facebook or Twitter. It excludes blogs, which are addressed in Pennock's report, and also excludes trading and marketing sites like Amazon or eBay because these will be addressed in a forthcoming report on preserving transactional data. Though the report aims to provide a general overview that applies to a wide range of social media, the diversity of social media services makes it impossible to address them all. Noticeable exclusions include Snapchat[1] and Periscope[2] – the former because it adheres to a general policy of 'delete is our default'[3] that undermines systematic capture, and the latter because its terms of service[4] are largely similar to those of Twitter, which owns it. It will address solutions that use Application Programming Interface-, or API, based solutions to request streaming data directly from a platform within the constraints imposed by the platform's terms and conditions, or terms of service. In addition to the legal framework introduced by platforms, it also discusses the ethical implications of preserving user-generated data. The report examines obstacles to workflows and practice, including issues of selection and indexing posed by the interlinked, conversational nature of social media. While the report offers a number of solutions for archiving social media, from user IDs to third-party services, it does not provide detailed guidance on any one approach. Rather, it points to other models and resources to support different approaches based on respective organizational needs.

Based on a study commissioned by the UK Data Service, the report will draw on use cases with data-driven research in order to inform appropriate recommendations and resources. It will further interrogate the use cases to provide relevant guidance for research and collecting institutions based on the needs of their user communities. These case studies will shape the report in order to articulate the current challenges to preserving social media as well as suggest possible solutions or alternative strategies. Though drawn largely from academic research institutions and cultural heritage institutions, the approaches demonstrated in the case studies have a much wider relevance to journalists and businesses.

Furthermore, while the report maps out a number of principle initiatives, it does not pretend to be an exhaustive survey of all current activities in the capture and preservation of social media. It will, however, provide a foundation from which to explore the wider arena of institutions currently archiving social media which may offer a relevant model for respective readers depending on the nature of their organization and the needs of its designated community. Table 1 below presents a list of the most popular social media platforms, focusing on user-generated and social networking platforms.

[1] https://www.snapchat.com
[2] https://www.periscope.tv
[3] https://support.snapchat.com/a/when-are-snaps-chats-deleted
[4] https://www.periscope.tv/tos

Table 1 ***Popular social media platforms 2014–15***[5]

Platform	Date Launched	Web URL	Function
Facebook	2005	facebook.com	social networking
Flickr (Yahoo)	2004 (bought by Yahoo 2005)	flickr.com	user-generated content
FourSquare	2009	foursquare.com	user-generated content, social networking
Google+	2011	plus.google.com	social networking
Instagram (Facebook)	2010 (bought by Facebook 2012)	instagram.com	user-generated content, social networking
LinkedIn	2003	linkedin.com	social networking (professional)
Pinterest	2010	pinterest.com	user-generated content, social networking
reddit	2005	reddit.com	bulletin board system, social networking
Twitter	2006	twitter.com	social networking
Vine (Twitter)	2012 (bought by Twitter 2012)	vine.co	user-generated content
YouTube (Google)	2005 (bought by Google 2006)	youtube.com	user-generated content

[5] These figures are in part based on statistics from we are social: http://wearesocial.com/uk/special-reports/global-statshot-august-2015. The other platforms reflect the use cases referenced in this report.

3. Introduction

This report addresses the challenges and solutions for preserving social media; in particular, it focuses on the concerns of archiving datasets in machine-readable format that will be of use on a large scale to data-driven researchers in the social sciences and other disciplines. The archiving of social media data as datasets, or as collections of big data, will allow research and collecting institutions to archive source data for academic research and journalism and preserve it for future access. It will also provide an important record with evidential value for legal and regulatory purposes. Overall, collections of social media data obtained through API-based strategies and adequately preserved by collecting institutions will help ensure the memory of an important cultural moment and help protect equal access to shared digital heritage.

For many collecting institutions, the preservation of social media has grown as an aspect of web archiving. The national libraries in the UK and Ireland have been building the UK Web Archive since 2004; this includes special collections that integrate social media content into the wider collection of mainly Web 1.0 sites (Pennock, 2013, p. 26). In 2014, for instance, the National Library of Scotland captured social media as part of their Scottish Independence Referendum collection[6] and the British Library recently captured social media around the 2015 UK General Election.[7] Because of the challenges of capturing social media using traditional methods such as web crawlers, the social media collections within web archives tend to be event-driven and limited to selected platforms, pages or user accounts. These strategies, however, have already been discussed in a previous *Technology Watch Report*, *Web-Archiving* by Maureen Pennock.

Archiving social media as datasets or as big data, however, faces different challenges and requires particular solutions for access, curation, and sharing that accommodate the particular curatorial, legal, and technical frameworks of large aggregates of machine-readable data. While Web 1.0 archives, typically stored as .warc files, can certainly be data-mined like social media, Web 2.0 social media platforms comprise machine-readable data generated by users in real-time, complicating the issues of capture and indexing even further than similar issues facing current web archiving. This report looks particularly at social networking sites and user-generated content sites, such as those listed in Table 1. The report excludes blogs because the methods discussed in *Web-Archiving* would be more appropriate for that medium. It also excludes trading and marketing platforms like Amazon or eBay because those types of services will be addressed in a forthcoming report on preserving transactional data. It also excludes live video or image sharing platforms like Snapchat and Periscope. Though both social media contain interesting, diverse, and culturally valuable content, this report does not have the scope to look in detail at strategies aimed particularly at these social media. Further research into the issues unique to these services is required, and should be pursued by institutions with a stake in the preservation of web content.

The Web 2.0 social media applications discussed in this report are comprised substantially of external links and varied embedded media types. Because these applications function to facilitate the exchange of information between users, often associated with fairly complex profiles, they also generate substantial amounts of secondary information, or metadata. The production of this type of metadata means that datasets can be potentially enriched with geolocation, date, time, gender, occupation, age, nationality, and a number of other derivative user elements (Sloan *et al*., 2015). The networked infrastructure of social media pages, as well as the underlying database systems that store user data, differentiate social media substantially from the traditional World Wide Web. Because of these qualities, social media requires a new set of definitions, solutions, and strategies for effective preservation.

3.1. What is social media?

In her report 'Archiving Social Media in the Context of Non-print Legal Deposit', Helen Hockx-Yu defines social media as: 'the collective name given to Internet-based or mobile applications which allow users to form

[6] http://www.nls.uk/collections/topics/referendum
[7] http://britishlibrary.typepad.co.uk/webarchive/2015/03/2015-uk-general-election-web-archive-special-collections.html

online networks or communities' (2014, p. 2). The term 'social media' is used to refer to many different web-based platforms, from social networking sites (SNSs) such as Facebook, or Qzone in China, to user-generated content sites (UGCs) like YouTube or Wikipedia. It is also used in some contexts to apply to trading and marketing sites (TMSs) which focus directly on selling, buying, trading, and advertising, including Amazon and eBay (van Dijck, 2013, p. 8). These different platforms – accessible through different types of application and on different types of device – possess a wide variety of functions and appeal to different audiences. From Twitter to YouTube, however, they all create a by-product of valuable data about the users who interact with them. Though this report will draw on use cases from the more popular platforms for data-mining among researchers and other users, such as Twitter and Facebook, it will generally apply to Web 2.0 applications that capture user-generated data in machine-readable languages that, in many cases, is monetized by commercial platforms and third party resellers.

Though social media is often synonymous with Web 2.0, social networking sites actually originated in early versions of the Internet, as far back as ARPANET.[8] In 1979, a programme called Usenet was developed to host news and discussion groups (Banks, 2008). This early form of social networking, which evolved into Bulletin Board Systems and then into sites like Friendster and MySpace, has summarily been replaced by newer generations of Internet-based networking sites.[9] Usenet, however, represents many of the underlying structures of Web 2.0: an interactive forum that facilitates the exchange of information between individuals, many of them strangers. As Internet availability increased in the late 90s and early 00s, however, another important change occurred that distinguishes Web 2.0 social networking. As social media applications grew in popularity, platforms, their capital investors, and their resellers developed a business model for monetizing information generated by users, particularly by engaging with the commercial sector looking for consumer analysis and market research. In the last decade, social media platforms have become competitive, vying both for users and for clients to purchase the valuable data produced by the interactions of its users.

This business model of monetizing user data has also increased the efficiency of capturing and encoding social media content; in other words, the data itself comes in a form easily processed by computers with enough metadata to derive significant information about users and their behaviour. Though this approach makes it easier to sell data to corporations, it also enables non-commercial researchers to access data for academic studies. With access to machine-readable data, researchers are able to process large samples relatively quickly. From studies into international response after natural disasters to the spread of hate speech, to prediction of election outcomes, social media provides a source of data that can detect patterns at an unprecedented scale through computer processing. More broadly, large-scale social media data has also been recognized as a significant cultural artefact. In 2010, the Library of Congress received a gift of all of Twitter's archived tweets from 2006–2010 and all on-going tweets. When the Library of Congress accepted the gift, Librarian James H. Billington emphasized the importance of social media in modern archival collections: 'the Twitter digital archive has extraordinary potential for research into our contemporary way of life. … Anyone who wants to understand how an ever-broadening public is using social media to engage in an ongoing debate regarding social and cultural issues will have need of this material' (Library of Congress, 2010). As Billington asserts, social media data provides a valuable record of contemporary life and offers important opportunities for understanding human behaviour and interaction.

3.2. Who archives social media?

While the commercial sector continues to improve its methods for gaining access to social media user data for consumer analysis, the academic sphere has also begun to develop methodologies for using social media as source data for studies in the social sciences, computer science, and a number of other disciplines. The UK Data Forum has integrated social media as a critical component of their 2013–2018 strategy:

[8] https://en.wikipedia.org/wiki/ARPANET
[9] http://www.digitaltrends.com/features/the-history-of-social-networking

> '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 (e.g. participation in activities; political attitudes; risk-taking behaviours; group motivations; religious beliefs; lifestyle choices, etc.). 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, p. 13).

In the past decade, the academic and archive sectors have increasingly recognized the value of social media data for research. These sectors, along with journalists and policy-makers, see the potential for social media research to improve public services and support better governance. As a result, these stakeholders have begun to devote resources to building methods and infrastructures for capturing and archiving social media data as big data, particularly academic social science. These new projects and initiatives, however, work within the strict boundaries of platform terms and conditions, ethical guidelines, and available resources for storage and technology. Several UK and EU research initiatives have begun to gather social media data in local repositories as a source for ongoing research at their institutions. In order to support this research, these initiatives have developed effective strategies for archiving social media datasets. The Social Data Science Lab at Cardiff University has begun using COSMOS, an integrated platform for studying social data, to capture 1% of all tweets daily through Twitter's streaming API, using the collection of data for a range of studies from the spread of hate speech to detecting cybercrime (Burnap *et al.*, 2014). The Social Repository of Ireland project led by INSIGHT @ NUI Galway partnered with Digital Repository Ireland has conducted a feasibility study to explore the potential for a repository of enriched social media data related to major events in Ireland.[10] These archives – formed as integral components of social science research initiatives – still face restrictions that limit the amount of data required to fully conduct relevant studies and even greater restrictions preventing the sharing of data, undermining a vital aspect of the research process.

These efforts in the academic sector mirror similar struggles in the archive and library sectors, where accessibility is even more highly restricted. As researchers creating social media datasets are prevented from sharing, the archives and repositories that traditionally support them are limited to storing either nothing of the social media data harvested by researchers or an extremely limited amount of metadata about those datasets, such as user IDs or, in the case of Twitter, tweet IDs. GESIS Leibniz Institute for the Social Sciences, for example, stores researchers' social media datasets, but only as tweet IDs (Kaczmirek and Mayr, 2015). Archiving social media as part of the historical and cultural record faces even greater challenges, not least of which the common platform restrictions on sharing. Despite limitations, however, some archives have successfully captured and curated social media archives available to the public. The UK Government Web Archive's social media collections at The National Archives capture official government Twitter accounts and YouTube channels through platform APIs and make the data available in JSON (JavaScript Object Notation) and XML publically on their website.[11] While there are not many examples of open social media archives in machine-readable format as part of heritage collections, the ongoing development of strategies for archiving social media as big data lays the groundwork for these institutions to do so in the future.

Though these initiatives have made inroads into the practice of archiving social media and maintaining it for long-term access, researchers, archives and libraries still face considerable challenges to preserving social media. Access to this data in machine-readable formats is still limited and often requires a specialized skillset, creating an inequality amongst the research community. The amount of data available through streaming APIs is limited, if available at all, and purchasing data from resellers is expensive and entails restrictive terms Curating and indexing social media data on a large scale as well as affording the needed storage space pose issues for making this data accessible to users. The legal and ethical dilemmas stemming from the large

[10] http://www.irishtimes.com/news/science/all-the-data-that-s-fit-to-print-and-archive-for-posterity-1.1965955
[11] http://www.nationalarchives.gov.uk/webarchive/twitter.htm and http://www.nationalarchives.gov.uk/webarchive/videos.htm

amounts of personal data within social media further complicate potential strategies for long-term access. Overall, the novelty of social media data paired with the particular economic and legal context within which it has developed make the issue of archiving social media a difficult task for all institutions looking to preserve this new and valuable source of social data.

3.3. How is social media archived?

As opposed to the static webpages, often referred to as Web 1.0, which can be harvested by web crawlers such as Heritrix, Web 2.0 content, like social media platforms, is more effectively archived through APIs. APIs are more effective than crawlers because of the fundamental differences in how Web 2.0 web applications operate. The boundary between 'Web 1.0' and 'Web 2.0' is not a definitive one, however; many attributes typical of Web 2.0 have particular implications for archiving and maintaining long-term access. In *Web-Archiving,* Pennock defines Web 2.0 as 'a term coined to refer to interactive, social and collaborative web technologies, resulting in a more distinctive and more modern World Wide Web' and emphasizes that much of Web 2.0 is 'commonly rich in JavaScript applications which … can cause problems for crawlers' (2013, pp. 36, 12). Tim O'Reilly originally coined the term to describe the shift from the web as a medium for publishing information (Web 1.0), to the web as a medium for participation (Web 2.0) (quoted in Helmond, 2015, p.5). Anne Helmond in 'The Web as Platform' furthers this distinction by demonstrating how 'an important aspect of Web 2.0 services are their software infrastructures for capturing, storing, organizing and redistributing data' (p. 6). Helmond argues that the interoperability and decentralization created by the design of social media platforms – particularly the ability to share core functionality such as 'likes' and 'comments' – has led to 'the "platformization" of the web' (p. 7). This 'platformization', created through APIs and other methods for sharing data, such as RSS and widgets, helps to explain the need for new approaches to archiving social media; in other words, the need to capture the residual information generated through what are essentially web receptacles. As opposed to snapshots, archiving the underlying data of social media applications preserves a more authentic and complete record.

An API is a kind of back door into a social media platform, though they are used for other types of dynamic websites as well. APIs allow developers to call raw data directly from the platform, including content and metadata all transferred together in formats like JSON or XML. The amount of data available through an API varies from platform to platform and many platforms have several different types of APIs. Twitter provides a Streaming API that gives 'low latency access to Twitter's global stream of Tweet data'.[12] In the context of harvesting data for analysis, this API creates a back door into current activity on Twitter but does not allow for searches of historical tweets. Twitter provides different volumes of streaming data, from 1% of tweets to the Firehose that opens access to 100% of tweets, though the Firehose requires special permission to access.[13] Facebook provides more limited access through APIs, in part because most user profiles are private and users are able to restrict the visibility of their content. Facebook, therefore, does not provide a Streaming API similar to that of Twitter, but instead offers a Graph API, or 'Social Graph' API, which gives access to the connections between users and also content shared on public pages.[14] These are just two examples of APIs. To access data through an API, it would be best to first explore the developer pages of a particular platform.

Research and collecting institutions may also obtain social media data from an API by licensing the data from a third party reseller. A reseller can supply API data alongside other services. Twitter's official reseller Gnip, owned by Twitter since 2014, provides a range of historical APIs including a Full-Archive Search API and the Historical PowerTrack that supply clients with raw data and data enrichment services (Messerschmidt, 2014).[15] A number of resellers provide licences to Facebook data, but DataSift is one of the biggest and the

[12] https://dev.twitter.com/streaming/overview
[13] https://dev.twitter.com/streaming/firehose
[14] https://developers.facebook.com/docs/graph-api/overview
[15] https://gnip.com

only to offer Facebook Topic Data.[16] Licensing data through a reseller is typically expensive, priced for corporations who want to analyse social media data to improve profits. In the past, Twitter's Gnip has advertised their support for research (follow URL in footnote to see a snapshot of the website in 2014), but primarily, along with most data resellers, targets for-profit corporations.[17]

In a few cases, research and collecting institutions have entered into direct agreements with platforms to obtain social data. The Library of Congress entered into, perhaps, the most prominent example of this type of agreement in 2010 when they accepted a donation of Twitter's entire archive and all on-going tweets (Library of Congress, 2013). In a different type of agreement with the Massachusetts Institute of Technology (MIT) called the Laboratory for Social Machines (LSM), Twitter has provided access to the entire Twitter archive and all ongoing tweets as well as committed US$10 million (£6.5 million) to a project.[18] Though these agreements with Twitter show willingness by the platform to make user data more accessible for non-commercial use, they do not necessarily offer a conclusive solution to the problem of accessing data for research use or for heritage collections. The Library of Congress has not (at the time of the publication of this report) provided access to the Twitter archive and MIT has not demonstrated that they will implement any measures to preserve the research data used in the project. Furthermore, these ad hoc arrangements create unequal access to data, giving advantage to well-funded institutions or those who are able to form a special relationship with a commercial platform.

3.4. What are the challenges to preserving social media?

Maintaining long-term access to social media data faces a number of challenges and varies from institution to institution. For researchers and collecting institutions alike, however, access to social media data poses a significant difficulty. Accessing data through APIs provides the most authentic record of social media, but developer policies and agreements attached to APIs restrict sharing and re-use. These contractual challenges facing the use of APIs for research and heritage institutions apply to any Web 2.0 content. The challenge of preserving social media data, in particular, entails the further difficulty of working with user-generated content that could contain identifying attributes to individuals. Researchers and collecting institutions must consider the risks to privacy and data protection introduced through processing large amounts of user-generated data. In planning long-term preservation, collecting institutions must assess the potential of these risks when accessing contemporary data in the future. For collecting institutions, the privacy issues surrounding user data further complicate already complex requirements for selecting and indexing social media content for re-use.

Despite these concerns, researchers and collecting institutions have begun to develop solutions for archiving social media for non-commercial purposes within the constraints of platform terms and conditions. For the majority of research and collecting institutions, the primary obstacle to archiving social media is accessing, or acquiring, the data. The following section outlines possible approaches to obtaining social media data for long-term access and the implications of those strategies for digital preservation.

[16] http://DataSift.com/products/pylon-for-facebook-topic-data
[17] http://web.archive.org/web/20141103063350/http://gnip.com/industries/research
[18] http://socialmachines.media.mit.edu

4. Strategies

Researchers and collecting institutions have developed multiple approaches to capturing and archiving social media from APIs as curated datasets or as big data. As discussed in the introduction, the principle strategies include harvesting data directly from platform APIs, licensing API data from a third-party reseller, and in rare cases, negotiating an agreement directly with a commercial platform. However, institutions have employed these strategies in different ways in order to make the most of their resources and to best meet the needs of their user communities. Collecting social media data is a new practice, therefore most institutions capture, curate, and store their data based on processes developed through experience with analogous types of content, such as web archives, research data, or content containing personal or sensitive data. Some best practices and standards for archiving and data management may provide guidance, but the extreme velocity and volume at which social media data grows exceeds the capacity of most established standards. The external restrictions imposed by terms and conditions, often by the terms of service and specifically by developer policies, further undermine the guidance of most standards for digital content management, for example the increasing requirement for open data by research funders. New best practices, however, can be observed in a number of successful initiatives, some of which will be discussed in the case studies later in this report.

4.1. Application Programming Interfaces (APIs)

Application Programming Interfaces are provided by social media platforms to enable controlled access to their underlying functions and data. Wikipedia defines an API as follows:

> 'An API expresses a software component in terms of its operations, inputs, outputs, and underlying types. An API defines functionalities that are independent of their respective implementations, which allows definitions and implementations to vary without compromising the interface. An … API can also assist otherwise distinct applications with sharing data, which can help to integrate and enhance the functionalities of the applications.
>
> APIs often come in the form of a library that includes specifications for routines, data structures, object classes, and variables. In other cases, notably SOAP and REST services, an API is simply a specification of remote calls exposed to the API consumers'.[19]

In the context of social media, an API acts as an interface between the social media platform and a consumer of social media data. The API defines how the consumer can interact with the platform in technical terms, and may define rules and restrictions on the access provided. For example, the Timehop application shows a user's past Facebook activity on a particular day by extracting data from the Facebook platform via the Facebook API.[20] Researchers or collecting institutions can also use APIs to pull raw data from social media platforms for computational analysis. In the work by Anne Helmond referenced in Section 3.3, Helmond engages with the description of social media platforms as 'walled gardens' or closed information systems. She argues that APIs 'enable these so-called walled gardens to plant their seeds outside of their gardens' (Helmond, 2015, p. 8). Just as developers use APIs to build new applications, researchers and collecting institutions can acquire social media data through APIs that act like doors to the otherwise closed world of social media platforms.

[19] https://en.wikipedia.org/wiki/Application_programming_interface
[20] http://timehop.com

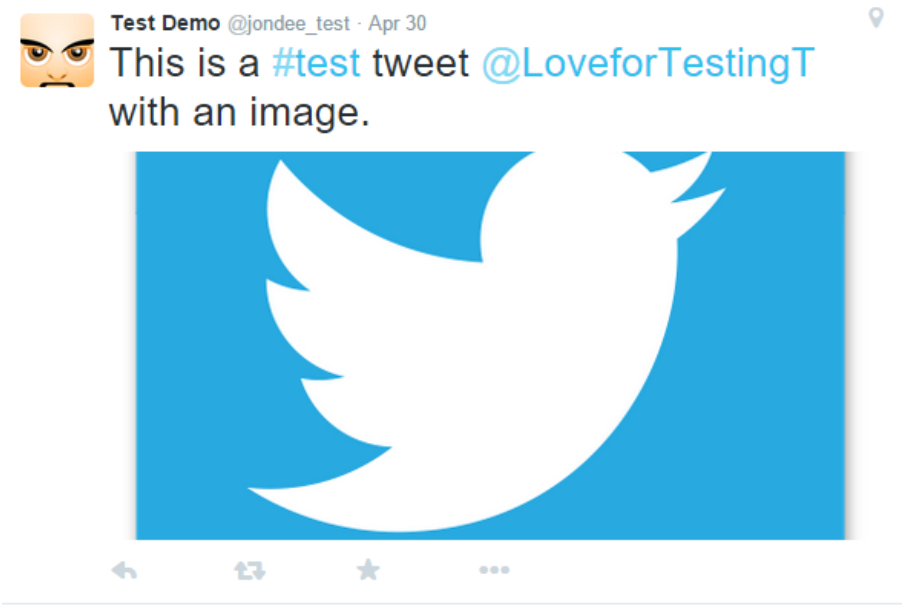

Figure 1: Rendered tweet from Gnip 'Data Format, Sample Payloads'[21]


```
{
    "created_at": "Thu Apr 30 21:53:11 +0000 2015",
    "id": 593895901623496700,
    "id_str": "593895901623496704",
    "text": "This is a #test tweet @LoveforTestingT with an image. http://t.co/ZvgHovKZq4",
    "source": "<a href=\"http://twitter.com\" rel=\"nofollow\">Twitter Web Client</a>",
    "truncated": false,
    "in_reply_to_status_id": null,
    "in_reply_to_status_id_str": null,
    "in_reply_to_user_id": null,
    "in_reply_to_user_id_str": null,
    "in_reply_to_screen_name": null,
    "user": {
        "id": 2993982541,
        "id_str": "2993982541",
        "name": "Test Demo",
        "screen_name": "jondee_test",
        "location": "Denver, CO",
        "url": null,
        "description": "this is a test account.",
        "protected": false,
        "verified": false,
        "followers_count": 2,
        "friends_count": 43,
        "listed_count": 0,
        "favourites_count": 0,
```


Figure 2: JSON of tweet in Figure 1 (lines 1–26 of 201) from Gnip 'Data Format'[22]

[21] http://support.gnip.com/sources/twitter/data_format.html#SamplePayloads
[22] http://support.gnip.com/sources/twitter/data_format.html#SamplePayloads

APIs can provide access to a stream of raw social media data, perhaps as it is created on the platform by users. The Twitter streaming API allows researchers and collecting institutions to obtain tweets generated by users in real time. Due to the massive volumes of data generated, social media platforms restrict the amount of data that can be requested via the API. In the case of Twitter, a developer or other user would apply for access tokens to 'spend' on a request for data.[23] If the access token is accepted, then Twitter will provide access, or 'open the door' to its streaming data. The streaming API provides ongoing data created from the time a developer connects to the API[24] and the requested data is transferred to the developer in a structured JSON format (see **Figure *1*: Rendered tweet from Gnip 'Data Format, Sample Payloads'**and **Figure *2*: JSON of tweet in Figure 1 (lines 1–26 of 201) from Gnip 'Data Format'**). Twitter's 1% streaming API provides a significant amount of data for most academic studies; however, Twitter does not disclose how the 1% sample is selected, preventing researchers from verifying if the data contains bias.

The JSON Twitter data provides the content of tweets along with elements of secondary information such as user ID, geolocation, and actions performed on the post, such as shares or likes. Twitter alerts developers that 'the attributes of a JSON-encoded object are unordered' and not to 'rely on fields appearing in any given order'.[25] While APIs differ from platform to platform, the underlying function of an API as a door into the closed system remains the same. This access provides a useful mechanism for obtaining social media data, but also comes with a number of restrictions. As described in detail later in this report, many APIs are governed by developer policies and agreements. Any content obtained through an API must further adhere to a platform's terms of service, including the user agreement.

APIs can also provide access to specific data, perhaps generated by particular users. Helmond, for instance, describes a technique of using the Facebook Graph API to replace components missed by web crawlers, such as Facebook comments displayed on a news outlet page (Helmond, 2015, pp. 134–144). The UK Government Twitter video archive uses the YouTube API to archive official government YouTube feeds but also to obtain the metadata associated with the videos. In addition to providing the primary content for a dataset or archive, APIs can support other archived media, such as web archives or digital corporate archives.

4.2. Data Resellers

Data resellers are companies that provide products and services based on data harvested from APIs. Some resellers, like Gnip, are the official reseller for a particular social media platform and provide access to some API data not accessible directly from the platform itself, such as historical data. As discussed in the introduction, data resellers provide tools and services that support the analysis of social data for market and consumer research. In 2011, Gnip released a PowerTrack service that offers access to 100% of tweets, costs $2,000 per month, plus $0.10 per 1,000 tweets delivered (follow URL in footnote to see a snapshot of the website in 2014).[26] In the past, they have promoted their support of research through services needed by researchers, particularly historical data services (follow URL in footnote to see a snapshot of the website in 2014).[27] Unlike Gnip, DataSift does not promote any services aimed at researchers, but does offer a wide range of services with access to a large number of data sources. Most notably, DataSift is now the exclusive provider of Facebook Topic Data which supplies 'anonymous and aggregated content data on what audiences are sharing on Facebook'.[28] For Twitter data, when researchers request data containing particular keywords or hashtags, as long as the targeted content does not exceed 1% of all tweets, they can harvest all relevant data. Researchers, therefore, do not often need to purchase data from resellers. These reseller services,

[23] https://dev.twitter.com/streaming/overview/connecting
[24] https://dev.twitter.com/streaming/overview
[25] https://dev.twitter.com/streaming/overview/processing
[26] http://web.archive.org/web/20150808184017/https://gnip.com/company/news/press-releases/announcing-powertrack
[27] http://web.archive.org/web/20141103063350/http://gnip.com/industries/research
[28] http://DataSift.com/products/pylon-for-facebook-topic-data

however, often exclusively provide access to historical data and 100% of streaming data that could be crucial to heritage institutions with a remit to preserve complete historical cultural records or artefacts.

The data sold by resellers, however, comes from platform APIs and therefore remains under the regulation of platform policies. In other words, clients may be able to obtain data from a reseller, but they do not acquire the right to publish or share that data. The target clients for data resellers are businesses, not research or collecting institutions. Businesses are primarily concerned with current data in order to monitor their current performance and improve sales (Puschmann and Burgess, 2014, p. 48). As a result, historical data is less of a priority for resellers. While Gnip and Datasift both offer access to historical data, their main focus is on products and services that draw on more recent data. This business model could make it difficult for researchers and collecting institutions to acquire access to historical data, which is more useful for longitudinal studies or heritage collections than for Fortune 500 companies.

4.3. Third-party services

Acquiring social media data through a platform API requires a degree of specialized knowledge of software development, and purchasing data from a reseller may incur high costs. As an alternative, particularly in the government and heritage sectors, collecting institutions could collaborate with a third-party archiving service. Several organizations provide social media archiving as part of their web archiving services or as their primary service. While there are a number of commercial services that specialize in archiving social media content, such as ArchiveSocial,[29] MirrorWeb,[30] Erado,[31] and Gwava,[32] other organizations such as the Internet Memory Foundation (IMF)[33] and the International Internet Preservation Consortium (IIPC)[34] also provide support for managing the archiving of social media. ArchiveSocial offers services based on number of records per month, identifying any instance of social networking such as an individual tweet or status update as a 'record'.[35] Service providers can offer custom services catered to the needs of a client, based on their size, user needs, or required frequency of harvests. Many of the commercial services that provide archiving for social media emphasize their usefulness for government organizations and companies with compliance needs and, therefore, generally offer competitive support for long-term access. Though third-party services can be costly, the development of an infrastructure for archiving social media can also entail high costs, including the cost of hiring staff with relevant expertise.

4.4. Platform self-archiving services

Self-archiving describes a back-up service provided by a handful of social media platforms that allows users to download the data from their account. A few popular platforms provide this function in user settings and allow account owners to download their content in machine-readable formats. Facebook, Google, and Twitter, for example, all provide this back-up service to different degrees (Bandziulis, 2014). Facebook will archive only content uploaded by the account owner or sent directly to the account owner. Once requested, Facebook send, via the associated email address, a .zip file of structured data (including photos, status updates, and messages) that is available for download within a few days, then expires. Twitter provides a similar archiving service for the tweets published from an individual user account. Twitter's self-archiving service comes with a navigable archive of tweets as well as JSON and CVS data for data mining (Bandziulis, 2014). Google provides self-archiving for each of their different services, including Gmail, Google Calendar, Hangouts, and YouTube. While these archives for individual user accounts are limited to the content held within those accounts, self-archiving may be a useful option for institutions that want to preserve their own

[29] http://archivesocial.com
[30] https://www.mirror-web.com
[31] https://www.erado.com
[32] http://www.gwava.eu/en
[33] http://internetmemory.org/en
[34] http://www.netpreserve.org
[35] http://archivesocial.com/pricing

institutional social media accounts. It could also be a solution for authors, politicians, or other public figures when depositing their personal archives. More generally, the personal archive of an institution's social media could supplement other web archives or digital archives or act as an index for other digital activities occurring at a particular time. Excellent guidance on how to use these self-archiving functions can be found through Wired[36] and the NCSU Libraries Social Media Archive Toolkit.[37]

4.5. Research dataset identifiers

After a researcher or research team has created a dataset, it is not usually possible for them to deposit that dataset with an archive or collecting institution for re-use. In the case of Twitter, one of the most popular data sources among researchers, the Developer Agreement & Policy restricts developers from sharing any data they obtain from the API and also from storing data in a cloud (Twitter, 2015).[38] The policy does, however, allow the archiving of tweet IDs, the unique number given to an individual tweet, or user IDs, the number assigned to Twitter account holders (*ibid.*). Other researchers could use the tweet IDs to recreate a dataset used in a previous study, but only if Twitter continues to provide access to historical data through Gnip. The identical recreation of a dataset is also only possible if no tweets are deleted after the original dataset is created (Summers, 2015). Despite these obstacles to recreating a dataset, archiving tweet or user IDs will provide a better solution than sharing no information at all about data sources for published studies. Furthermore, Twitter also allows the sharing of up to 50,000 tweets in a non-machine-readable format, such as a spreadsheet or PDF (*ibid.*). While this does not facilitate the re-creation of computational studies, it does provide a solution for developing smaller collections of research data.

Researchers use different methods to access social media data from APIs – different tools, different platforms, different types of APIs, different resellers with different services – which create very diverse types of dataset. Furthermore, individual researchers use different methods to clean, or organize, their data, as well as different tools and methods for analysing their data. In addition to the IDs associated with a dataset, information about how the raw data was collected and how it was 'cleaned' is also important and will be required for re-creating a dataset or understanding how and why it has been altered (Weller and Kinder-Kurlanda, 2015, p. 34). Therefore, the archiving of dataset identifiers is more effective if the processes used to create them are also documented.

4.6. Research documentation and metadata

One of the primary reasons why it is important for researchers to share their source data is to ensure other studies can reproduce and validate their results. The ability of peers to recreate a particular study with the same methods on the same data substantiates findings and strengthens the science behind innovative ideas and theories. In the past decade, more and more of these ideas and theories have derived from social media data. However, due to the novelty of social media data, standards and methods are not firmly established across disciplines. Some fields, though, have begun to develop sophisticated and tested strategies for working with social media data, particularly in the social sciences. As computational social science increasingly draws on social media data, the documentation of workflows and actions taken on the data will support short-term preservation when source data cannot be shared.

As with the sharing of tweet IDs, the standardization and sharing of social media research data does not provide a complete solution, but it may prevent the complete loss of important information about how data is obtained and processed. In a recent study looking into the practices of social media researchers, participants revealed varying methods for documenting their work and expressed the difficulty of such an undertaking, particularly with no standard guidance to follow. As a solution, one participant suggested:

[36] http://www.wired.com/2014/07/archive-social-networks
[37] https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/facebook-and-twitter-personal-archives#note2
[38] https://dev.twitter.com/overview/terms/agreement-and-policy

> '… this task [of documentation] should therefore be outsourced to specialists such as librarians and archivists. Thus, the question remains of how to document work-flows and experiences in such a way as to make data and documentation accessible to the wider community, thus allowing sharing of background knowledge in order to enable comparative research and reproducibility of results' (Weller and Kinder-Kurlanda, 2015, p. 35).

As the research community develops methods for working with and documenting social media data, this metadata may provide the closest surrogate for data governed by commercial platform terms and conditions that prohibit sharing or depositing with a third party. Metadata containing information about workflows and process documentation is most effective for publishable research when it is standardized, and also deposited and made available to other researchers. Archives and libraries are, arguably, in the best position to provide this support with their history of expertise in creating and preserving metadata as a part of archived digital objects. If archives and libraries are to support the preservation of social media data documentation, researchers and archive and library professionals will need to collaborate closely. The Software Sustainability Institute (SSI) has begun to promote the collaboration of researchers and preservation experts to make software and processes available for the long term (Pawlik *et al.*, 2015). Work from SSI may provide a model for collecting institutions for developing support services for researchers.

5. Challenges

Twitter's Developer Agreement & Policy governs how data from Twitter's APIs can be used and altered. Under Section Two, 'Restrictions on Use of Licensed Materials', the Agreement forbids developers to: 'sell, rent, lease, sublicense, distribute, redistribute, syndicate, create derivative works of, assign or otherwise transfer or provide access to, in whole or in part, the Licensed Material to any third party except as expressly permitted herein'.[39] This clause prohibits the sharing of Twitter data access through one of the Twitter APIs and undermines the ability of researchers to deposit API data with a third party, such as an archive or library. Due to these restrictions, the outlook for sharing Twitter data is somewhat bleak for collecting institutions that support research or preserve culturally important content from this popular micro-blogging site. Furthermore, the 'Twitter API is more open and accessible compared to other social media platforms,' making the outlook arguably worse for archiving other social media platforms (Ahmed, 2015). Ultimately, objectives for archiving social media as big data for the purposes of cultural memory or non-commercial research are at odds with the business model of most social media platforms that depend on the monetization of their data for profit (Puschmann and Burgess, 2014).

Though the restrictions imposed by commercial platforms present the primary challenge for most institutions, the effort to save this valuable content faces a multitude of other potential hurdles. After the difficulty of *obtaining* social media data, ensuring privacy and security for individual user information also poses difficulties. The application of computational analysis to large sets of data increases the risk of accidental disclosure, making it important to ensure archived data have had any trace of identifiable data removed (Executive Office of the President, 2014). The long-term preservation of user-generated content is also at odds with emerging EU legislation to protect the right to be forgotten. EU rulings on an individual's right to have their personal information removed from Internet search engines in certain circumstances has a significant impact on the practices of organizations working with digital content sourced from the web (Koops, 2011). In terms of techniques and workflows, social media data pose a particular challenge for curation and selection, due to the volume and complexity of the data. Furthermore, researchers and collecting institutions have no control over the algorithm used to sample data from APIs. Without knowledge of how data is sampled, or selected, research and collecting institutions are unable to prevent bias from entering the archive or research study. And once a dataset is obtained, researchers and collecting institutions will face a difficult task in storing and indexing the data in a way that will make it useable for future users. These challenges include finding methods for preserving user experience that will convey the context that social media interactions take place in.

This chapter outlines the overarching challenges to preserving social media posed both by external concerns and those inherent to the nature of the data. While this report acknowledges that all new forms of digital content create a need to re-evaluate current practices and methods, the external obstacles presented by both the commercial interests of social media platforms and the sensitive nature of user-generated content make this a unique problem. The following sections will address the key areas of concern when approaching the preservation of social media data.

5.1. Platform terms and conditions

As previously discussed, social media platforms are commercial companies that make a profit from selling user data, for instance to corporations for consumer analysis. In order to protect their profits, commercial platforms must ensure the security of their data, including contractual obligations to control what happens to their data once it is in the hands of a third party. One such protection is the API developer policy, which usually prohibits the sharing of data acquired through an API and imposes limits on how frequently data can be requested. These conditions, particularly the developer policy, reflect the underlying business model that relies on corporate data clients and also the development of applications that interact with Twitter. APIs, therefore, are not research friendly. When researchers or non-commercial collecting institutions access social

[39] https://dev.twitter.com/overview/terms/agreement-and-policy

media data, they face the same conditions as for-profit companies (including the Fortune 500 companies advertised as prominent clients of both DataSift and Gnip).

Researchers need substantial samples of data, depending on the parameters of a given study in order to derive significant patterns from social media data. However, most social media platforms that provide access to their API directly also restrict the amount of data that can be requested and how often, through rate limits. Social media platforms like Twitter track the requests made through their APIs in order to prevent excessive data access. (The repercussions for over-requesting could entail having privileges completely revoked.[40]) In addition to rate limits, social media platforms protect the algorithms used to generate the allowed sample size. Though Twitter, for instance, assures developers that the sample is completely random, without the algorithm used to generate the sample, researchers cannot verify that the sample does not contain any bias or misrepresentation. Researchers need to demonstrate their methods for creating a dataset when they publish their analysis. Social media researchers, therefore, struggle to publish their outputs for re-use in the academic community (Weller and Kinder-Kurlanda, 2015, p. 35). Rate limits may also restrict collecting institutions that want to build comprehensive collections about a city, a country, ongoing events or catastrophes, or other large-scale topics. Similarly, collecting institutions cannot prevent a sample bias from entering their heritage collections or public records.

For data-driven researchers, the ability to share data has become an increasingly important part of the research process and, in some circumstances, is required. The Digital Curation Centre, an organization that provides support for research data management for higher education research communities in the UK, stresses the importance of creating a plan for the management, sharing, and preservation of research data, particularly as funders increasingly require researchers to share this.[41] Though some social media platforms make allowances for accessing data used for research, they do not extend to sharing datasets openly in digital repositories. Twitter, for instance, allows researchers to provide the tweet IDs for the tweets in a particular dataset, thereby allowing other researchers to request the same set of tweets as the original. The Developer Agreement & Policy states: 'If you provide Content to third parties, including downloadable datasets of Content or an API that returns Content, you will only distribute or allow download of tweet IDs and/or User IDs' (Twitter, 2015). While this provides a means for researchers to attempt to validate earlier results, it does not ensure accurate outcomes. In many cases, a cloned request of public tweets will not yield the same dataset. For instance, due to user action – such as deletion or editing of tweets – a dataset could contain different content. Despite exceptions like that of Twitter, these restrictions prevent researchers and collecting institutions from sharing their data in a meaningful way. For public collecting institutions, with a remit to preserve national heritage, the inability to share social media collections undermines their essential purpose.

The terms and conditions that govern social media data also restrict how that data can be stored. The section of the Twitter Developer Agreement & Policy quoted above that forbids developers to 'sell, rent, lease, sublicense ...' etc., indirectly influences how researchers and collecting institutions move data around once they have acquired it.[42] Strict adherence to this clause would prohibit storing any acquired Twitter data in cloud storage as this would involve transferring the data to a third party storage provider. Twitter is not unique in licensing data under a non-transferrable agreement – Foursquare, LinkedIn, and YouTube all uphold similar licensing restrictions.[43] Unfortunately, many organizations facing long-term storage needs without the budget to build a local storage facility may find a solution in cloud storage.

Platform terms and conditions limit how researchers and collecting institutions can harvest and preserve social media. Furthermore, these terms and conditions change fairly frequently, sometimes every year or few

[40] https://dev.twitter.com/overview/terms/agreement-and-policy
[41] http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies
[42] https://dev.twitter.com/overview/terms/agreement-and-policy
[43] Foursquare: https://foursquare.com/legal/api/platformpolicy; LinkedIn: https://developer.linkedin.com/legal/api-terms-of-use; YouTube (Google): https://developers.google.com/youtube/terms?hl=en

years (Lanigan, 2015). This makes it difficult to establish long-term policies for handling the licence restrictions on social media archives. These variables will impose a greater challenge for researchers or collecting institutions that collect content from multiple platforms on a regular basis. Frequent review of the relevant social media terms and conditions, though time-consuming, could ensure compliance with social media platforms. In addition, platform terms and licence conditions constitute only one factor, albeit the dominant factor, in the legal and ethical contexts that influence how social media can be archived.

5.2. Copyright infringement

Terms and conditions often pre-empt copyright infringement issues in the UK, and in countries with similar legislation, because they restrict acts of copying and distributing API data. Furthermore, the harvesting and processing of data does not necessarily involve copying or re-distributing protected material, even if some material published on social media does enjoy copyright protection. For instance, in the *Morel* v. *Agence France-Press* case, US courts ruled that photographs posted on social media sites are protected by copyright.[44] Based on this decision, a Twitter dataset may contain proprietary images, but analytics performed on that dataset may only query the metadata related to the post and not involve copying or sharing of those images. Furthermore, any resulting academic publications will likely only print high-level quantitative outputs but not re-display the original images. The licensee is prohibited by Twitter's developer policy from sharing copies of raw data – and thereby infringing copyright – therefore copyright infringement is unlikely to be a risk in the long-term preservation of social media data under current conditions.

Regarding the possibility of infringement when making a preservation copy of such a dataset, Twitter's user agreement stipulates that Twitter owns the right to share all data published publically by its users and the Twitter Developer Policy governing the API does not limit the number of copies kept by the licensing party (Twitter, 2015). While the ethical quandary of archiving content created by private individuals who may not be aware of this policy poses potential problems, the issue of copyright infringement in this scenario does not present a significant risk. Copyright issues may be more pertinent when it comes to qualitative analysis of social media because authors may want to reproduce individual tweets or social media content in their publications. However, the infringement of copyright in the instance of researchers using individual pieces of copyright-protected content, such as images or sound, does not become an issue unless the researcher wants to publish or otherwise distribute a copy of the image contained in the data. If researchers publish analysis of large aggregates of user data (as opposed to individual user accounts or posts), there is less (or no) risk of copyright infringement. Therefore archiving and preserving large aggregates of user data from social media APIs poses very little risk of infringing copyright.

5.3. Brokering agreements with platforms

To date, only Twitter has actively deposited user data with a trustworthy archival institution. It has also brokered agreements with individual research institutions to support academic studies. Other than these, there is very little precedent for relationships between social media platforms and institutions who specialize in archiving and long-term preservation. Indeed, the lack of preservation planning by commercial web platforms has led to the loss of valuable user data. In some instances, archivists have been able to rescue data from closing web services, but emergency recovery is by no means a sustainable preservation approach. GeoCities, for example, was shut down in 2009 when Yahoo! 'succeeded in destroying the most amount of history in the shortest amount of time … in known memory. Millions of files, user accounts, all gone' (Archive Team, 2009). Today, users can access GeoCities only through the archives of rescued data, for instance those held by Archive Team (*ibid.*). Without a preservation plan or deposit agreement with an archival institution, social media records created on today's popular platforms may be lost or become un-interpretable in the future.

[44] NCSU Social Media Archives Toolkit: https://www.lib.ncsu.edu/social-media-archives-toolkit/legal

Currently, the business model that underpins commercial platforms and the data market does not provide for long-term preservation or long-term access for researchers or future user communities. In order to bridge this gap, research and collecting institutions and policy makers should develop an alternative business model which will reduce restrictions to non-commercial access without interfering with the profits made from corporate sales. Perhaps corporate sales could help fund non-commercial access, either by rewarding corporations who share their data with non-commercial institutions or by providing special benefits to those who make donations to research for data costs. It may also be feasible to consider using research infrastructure funding to establish a long-standing arrangement with willing platforms to transfer a certain amount of data for research or heritage collections. Regardless of any potential new business models, non-commercial data access does not need to interfere with commercial profits for user data. The following agreements demonstrate models for negotiating with Twitter and lessons learned from those experiences.

5.3.1. Twitter Research Access at the Library of Congress

The challenges imposed by terms and conditions largely stem from the efforts of social media platforms to protect profits gained from selling user data to commercial companies. These sales help them to afford the frequent re-architecting of their technology to accommodate the growth of users and activity on their platforms (Zimmer, 2015). Because of this profit strategy, large-scale agreements between commercial platforms and non-commercial entities are rare. One exception to this precedent is Twitter's 2010 gift of all archived and ongoing data to the Library of Congress. The Library refers to the process of archiving Twitter's gift as 'Twitter Research Access', but researchers may not actually have access to this archive any time soon. While granting the Library a donation of all tweets and accompanying metadata, the agreement with Twitter also established a number of conditions. These mostly govern access by the Library, including a six-month wait period after a tweet is published as well as access only for authorized ('bona fide') researchers (Library of Congress, 2013). While this gift from Twitter signals an intention to co-operate with the heritage sector and to support non-commercial research, the initiative has encountered some difficulties that reveal the specific challenges associated with such a collaboration (*ibid.*).

This gift will provide the Library of Congress with an important source of data for future researchers; however, the process of ingesting and organizing the incoming data has proved time-consuming. In 2010, the Library did not have the capacity to transfer data from Twitter, so this was outsourced to the third-party service Gnip, not yet acquired by Twitter (*ibid.*). The Library and Gnip have an established process for transferring the data, but the Library has yet to announce how the archive may be accessed. According to a white paper released in January 2013, the Library would have completed the organization of the 2006–2010 archive into hourly files by the end of that year (*ibid.*). However, because a simple full-text search of the archive took more than a day, they had not yet opened access to the archive, leaving more than 400 requests for research access denied (*ibid.*). Officially named the Twitter Research Access project, the Library has not communicated any updates on progress in making the archive available to researchers since 2013. In a summary about the status of the Library's Twitter Archive, Politico.com published a quote from Library spokeswoman Gayle Osterberg, who informed the news outlet that 'no date has been set for [the Twitter Archive] to be opened' (Scola, 2015). This statement reveals the extent of the challenges imposed by a wholesale agreement with a social media platform.

In 2013, the Library attributed the delay in opening the archive to the difficulties of organizing and indexing the data, engineering a method of access, and to data protection and privacy issues surrounding content generated by public users. Any specific or updated information about individual challenges or methods for solving them has not been communicated. More general information, however, about the institutional shortcomings of the Library since accepting the gift from Twitter was revealed in a report from the US Government Accountability Office (GAO) in March 2015.[45] This report focused on the Library's information technology management overall, but specifically cited the Twitter Research Access project as a case study for

[45] http://www.gao.gov/assets/670/669367.pdf

leadership failings within the Library's management structure. In particular, the report found that delays in delivering Twitter Research Access are a direct result of the failure of the Library to observe institutional procedures when taking on the Twitter archive, such as bypassing the Library's selection process, approving the project through the Executive Committee instead. According to the GAO report, the failure to process the Twitter donation through official selection procedures has led to the lack of a schedule, cost analysis, scope, and strategic impact at the 'enterprise level' (US Government Accountability Office, 2015). In addition to an overall absence of oversight for Library IT investments, this lack of planning has contributed to the inability of the Library to open the Twitter archive to researchers.

The agreement between Twitter and the Library of Congress represents an important step in the preservation of social media. Putting the agreement into practice, however, has been slowed by the unprecedented challenges of archiving this novel form of digital content, entailing technological, curatorial, legal, and ethical challenges beyond those associated with other forms of digital content. Despite the shortcoming reported by the US GAO, the Library of Congress has an established infrastructure for the ingestion and preservation of digital content and successfully manages other digital collections; this emphasizes the importance of planning and selection to archiving such a high volume of social media data. The problems highlighted by the continued lack of access and the GAO report indicate some specific challenges to an agreement that transfers all archived and ongoing content at such a scale. As the website *Inside Higher Education* puts it: 'the Library of Congress finds itself in the position of someone who has agreed to store the Atlantic Ocean in his basement' (McLemee, 2015). This case study provides useful lessons for the research and archive communities, particularly on how best to form relationships with commercial social media platforms.

5.3.2. MIT's Laboratory for Social Machines

The agreement between Twitter and the Library of Congress is a useful case study because it aimed to ensure long-term preservation of Twitter user data. Agreements with research institutions – that do not involve long-term preservation – also provide a useful model for negotiating with the platform. In October 2014, Twitter invested US$10 million over five years into the MIT Media Lab's 'Laboratory for Social Machines' (LSM).[46] The agreement provides the LSM with access to streaming tweets as well as all historical tweets through Twitter's data service Gnip (Gillis, 2014). The LSM does not report any plans to ingest or archive the datasets they create from Twitter data. The agreement focuses on access to, rather than transfer of, data. The Laboratory of Social Machines (LSM) will use the data from Twitter to analyse the way cities and communities function through social media. Their research will support the development of tools and interfaces to further utilize social media as a means to make communities operate more effectively and make government more transparent (MIT News, 2014).

This type of arrangement with Twitter allows users from MIT to use social data for non-commercial research, but without the challenges of transferring all of Twitter's data to local facilities. The benefit of this is that the research institution avoids the high costs of archiving datasets. It does, however, reflect a unique relationship between Twitter and MIT. Associate Professor Deb Roy, who will lead the LSM at MIT, also serves as the chief media scientist at Twitter. Twitter and MIT have both benefitted from this relationship in the past, when Twitter purchased Roy's social television analytics company Bluefin, developed from MIT research, in February 2013 (Brustein, 2014). Though MIT has stated that '… the LSM will have complete operational and academic independence', the research will be performed at an organization that already enjoys close ties to the commercial platform (MIT News, 2014). While this type of agreement circumvents some of the practical and technical problems faced by the Library of Congress, the access granted to the LSM benefits the few over the many. Such ad hoc agreements may fail to realize the full potential of social media research, a scope that could be broadened if researchers across many institutions had similar access to Twitter data or were able to share datasets more freely. Deb Roy states that 'there are a lot of people at Twitter who are interested in leveraging Twitter for social good', and this investment at LSM signals an encouraging message that social

[46] http://news.mit.edu/2014/twitter-funds-mit-media-lab-program-1001

media platforms may play an increasing role in support of non-commercial access to social data (Brustein, 2014). The LSM and the relationship between MIT and Twitter will hopefully generate an important case study for collaboration between non-commercial institutions and commercial platforms, but as it stands, the LSM agreement provides a limited access solution, and does not answer the concerns of long-term access to Twitter social data for research and heritage more broadly.

5.4. Ethics of user privacy and awareness

Social media data, compared with other web content, pose particular concerns for long-term preservation because social media content is created by users – private individuals using social media services. Academic research, especially in the social sciences, has well-established methods and ethical guidelines for conducting analysis of materials that contain personal or sensitive information. In the social sciences, for instance, ethical standards are established through mechanisms such as the OECD's 'Guidelines on the Protection of Privacy and Transborder Flows of Personal Data'.[47] Ethical standards are also enforced through major funding bodies, such as the ESRC's Framework for Research Ethics.[48] In addition, researchers may look to media-specific guidelines for treatment of sensitive data, for instance the Association of Internet Researcher's ethical recommendations.[49] The archive sector too has produced mature standards and guidelines for the ethical treatment of digital content containing personal and sensitive data. In the UK, the Information Commissioner's Office has oversight for adherence to data protection and privacy issues in the information sector.[50] Questions of whether social media can be archived ethically are also reflected in EU case law regarding the right to be forgotten.[51] Social media data, however, presents new questions of ethics for archival preservation. In particular, when subjected to analysis, social media data may unintentionally reveal personal information about users, a risk that increases significantly when combined with data from other sources, such as administrative data. Furthermore, although much of social media is available publically on the web, users may not be aware of how their data is being used. Though users are the authors of social media content, many platforms own the right to transfer and sell that content without alerting them. This brings into question whether or not ticking a box when signing up for a social media account constitutes an acceptable indication of consent (Cate, 2012); the extent to which collecting institutions will be responsible for these issues of privacy and consent is yet to be established. Any institution which holds archived social media data, however, may inadvertently pose risks to private individuals. They may also undermine individuals' rights not to have data about them held by a third party for longer than the period in which the data is relevant or of public interest.

Researchers and policy makers do not yet fully understand the extent of the risk of accidental disclosure posed by the very large amount of information within big data. Therefore, taking measures now to store and secure 'big' social media data is even more important than for other forms of digital content (OECD, 2013, p. 23). The risk comes from the intrinsically linked nature of big digital data, which makes it easier to accidentally disclose the identities of private individuals. When multiple sets of data, from social media or administrative datasets for instance, are combined and subjected to analytics, connections may be made between individuals and their personal information. Though there are some methods to mitigate this risk, simple anonymization may not fully prevent such accidental disclosure. In a report to the Executive Office of the President in 2013, advisers on science and technology warned: 'Anonymization is increasingly easily defeated by the very techniques that are being developed for many legitimate applications of big data. In general, as the size and diversity of available data grows, the likelihood of being able to re-identify individuals … grows substantially' (Executive Office of the President, 2014, p. xi). These conditions make it imperative for

[47] Organisation for Economic Cooperation and Development (OECD): http://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm
[48] Economic and Social Research Council (ESRC): http://www.esrc.ac.uk/funding/guidance-for-applicants/research-ethics
[49] Association of Internet Researchers (AoIR): http://aoir.org/ethics
[50] Information Commissioner's Office (ICO): https://ico.org.uk
[51] http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=URISERV%3Al14012

researchers to adhere to new ethical standards and precautions when performing analysis on social media data, particularly when combining these data with other sources (Housley *et al.*, 2013, p. 175).

Cornelius Puschmann and Jean Burgess articulate the ethical issue of user awareness and control quite clearly in their article on 'The Politics of Twitter Data' (2014). In the article, the authors describe the shift of data ownership to commercial platforms and data resellers through highly technical access mechanisms like APIs and prohibitive developer policies.

> 'It follows that only corporate and government actors – who possess both the intellectual and financial resources to succeed in this race – can afford to participate, and that the emerging data market will be shaped to their interests. End users (both private individuals and non-profit institutions) are without a place in it, except in the role of passive producers of data' (Puschmann and Burgess, 2014, p. 52).

This ownership framework exists around more social media platforms than just Twitter. Facebook, Google, LinkedIn, and most other platforms all claim data ownership through their terms of service. This framework results in a system where end users have little control over what happens to posts and activity once they are published. Platform ownership of data, supported by corporate resellers, underlies the ecosystem that makes large amounts of user data available for commercial companies (Puschmann and Burgess, 2014). This ecosystem also makes large amounts of social data available for research, journalism, heritage collections, government records, and other non-commercial use; users are unlikely to know their data is being used for particular research analysis or included in archival collections. Because of this lack of awareness, some researchers have identified a conflict of interest in using user data or in disclosing direct quotations from user-generated content (Weller and Kinder-Kurlanda, 2015, pp. 33, 35).

Research and non-commercial collecting institutions have a different relationship with social media users than commercial companies do. Most analysis of social media data carried out by commercial companies focuses on recent data to increase sales and does not depend on the co-operation of users. By contrast, academic research and collecting institutions rely on the trust and assent of the community. Much of the funding for research and collecting institutions comes from public money through government budgets. It benefits these institutions, therefore, to demonstrate trustworthiness to the public – users of social media. More fundamentally, social media research that supports the public good – and also heritage collections maintained on behalf of the public – has an obligation to carry out these duties without doing harm to the public it serves. This obligation to avoid doing harm to individuals when saving their data over long periods of time is reflected in the principle of the right to be forgotten as established in Europe through the implementation of Article 12 of Direction 95/46/EC in multiple nations' case law (Mantelero, 2013, p. 232). With access to social media data, however, researchers are enabled to perform analysis that has the potential to greatly benefit society. Institutions who archive social media, therefore, must take measures to engender trust with that community. As the OECD Global Science Forum reported in February 2013, 'the challenge is to ensure that technology is used in ways that encourage the spread of best practice access arrangements, while ensuring that legal safeguards are adhered to and that the risk of inadvertent disclosure of identities is minimised' (OECD, 2013, p. 8).

Though commercial platforms have some measures in place to protect users, enforcement of user agreements often protects the value of data as much as the privacy of users. In summer 2015, Twitter rescinded API access from Politwoops, run by a non-commercial organization, the Sunlight Foundation.[52] Politwoops[53], based on a Dutch website with the same name[54], deliberately publishes (and publicizes) deleted tweets by politicians and official government accounts. In June 2015, Twitter revoked Politwoops'

[52] http://tktk.gawker.com/twitter-just-killed-politwoops-1708842376
[53] http://politwoops.sunlightfoundation.com
[54] http://politwoops.nl

access to the API citing a violation of the user agreement and developer policy that forbids the harvesting or preservation of deleted posts. Twitter chose to enforce the regulation, despite an agreement established in 2012. In this earlier agreement, Politwoops assured Twitter it would publish deleted tweets only for purposes of transparency and would use human controls to eliminate 'low-value tweets, links, and Twitter handles' (Gates, 2015). The non-profit also committed to tracking deleted posts only with the intention of 'revealing a more intimate perspective on our politicians and how they communicate with their constituents' and not to cause embarrassment or harm (*ibid.*). Considering Twitter honoured this agreement with Politwoops for three years without citing violations of user agreements, Twitter may have had other motives for rescinding Politwoops' API access. It may reflect the company's increasing interest in keeping its ownership of user data exclusive as much as it does with protecting the user agreements of US politicians. In other words, the decision could be about protecting profits as much as about protecting users. Though access was restored to Politwoops in December 2015, this action reflects the social media platform's change in attitude towards the use of user data generated by their service.[55]

As the use of social media grows and practices for archiving user-generated data mature and improve, the larger the ethical implications become. Based on the rate of growth in social media use, a dramatic increase in available data and expansion of social media research is likely to occur. The impact of this growth in social media prompted researcher Ralph Schroeder to characterize social media as 'a part of an essential infrastructure for citizens, along similar lines as broadcast, communication via phone, energy or transport' (2014, p. 9). As such, Schroeder argues:

> '… if these services are seen as essential infrastructures, it will be possible to argue that the large-scale manipulation of people's thoughts and feelings via online media could be deemed harmful. … combined, these [conditions] provide a more powerful picture, and thus a more manipulable population' (2014, p. 9).

Without regulation or intervention, the long-term effects of social media analytics – supported by archives of social media data – could lead to much more substantial repercussions than individual cases of accidental disclosure or lack of user awareness. Schroeder refers to these far-reaching consequences as 'creep', but the reality of a closed system where users have little or no control over their data could develop as rapidly as social media platforms gain users (2014, p. 9). By 2015, 29% of the world's population was active on social media, a 12% increase in only a year.[56] In 2016, more than a third of the world's population is likely to use social media, creating unprecedented amounts of social data.[57]

5.5. Selection

The practice of archiving social media requires methods to account for the linked, interactive nature of most platforms. In *Web-Archiving*, Pennock describes this issue of defining a collection of social media as opposed to traditional websites: 'Twitter, for example, is not just about tweets, but about the conversation. To archive a single tweet, therefore, is to archive only one side of the conversation. How do you establish the boundaries for a coherent Twitter collection?' (p. 13). Institutions have approached this question of boundaries in a number of ways. The social media archives created by the North Carolina States University's libraries, for example, focus on the capture of institutional Twitter and Instagram accounts based on official university account handles and university-related event hashtags. They also consider community responses that refer to these account handles and hashtags to be part of the official conversation and include this public Twitter and Instagram content in the collection (North Carolina State University, 2014–15). The Social Repository of Ireland project programmes its Twitter API harvesting tool to collect all tweets related to the Republic of Ireland, using geolocation, keywords, and hashtags (see Case Studies for more information). At academic repositories such as GESIS, social media is archived as the datasets created by researchers, thus

[55] http://www.nytimes.com/2016/01/01/technology/twitter-politwoops.html?_r=0
[56] http://wearesocial.sg/blog/2015/03/digital-social-mobile-in-apac-in-2015
[57] http://wearesocial.sg/blog/2015/03/digital-social-mobile-in-apac-in-2015

using the criteria of individual research projects to drive the nature of the social media in their holdings (see Case Studies). These approaches demonstrate attempts to define an archival collection of social media, but also highlight the challenges introduced by the new media. Each platform presents its own obstacles to selection and curation, but all platforms with a networking element, from Twitter to Instagram to YouTube, pose difficulties for deciding the content that comprises a coherent collection and, more pragmatically, the content that can be effectively preserved within the legal, ethical, and technical framework in which it is collected.

The 'conversation' of social media, particularly of social networking sites, makes it difficult to identify the boundaries of a collection and to establish selection criteria. The various pages and media that contribute to social media interactions extend across multiple user accounts and often evolve into related conversations or events without clear delineation of when one conversation ends and another begins. Identifying all the important entities (persons, places, institutions, event names, etc.) also poses difficulties, as users do not always use uniform hashtags or keywords and communicate in natural language that introduces variations in terminology and even spelling errors (Risse, Peters *et al.*, 2014, pp. 212–3). Establishing boundaries around a topic or major event or location provides a general scope for defining a collection. Limiting collection to a single platform or a small number of platforms may also make its scope more manageable. Even within these constraints, however, ensuring a complete collection in the bounds of these parameters has the potential to miss significant portions of the conversation. To some extent, the archiving of institutional or personal social media accounts circumvents this problem, such as at The National Archives' UK Government Twitter and YouTube archives that systematically exclude all responses and content published by any social media account outside official central government. However, this institutional approach must find alternative means of preserving the context of the conversation in order to ensure that future users will be able to interpret the archived tweets and videos (see Case Studies for more details).

For institutions that support researchers doing computational analytics, the importance shifts to capturing enough data to enable most research studies. Though the amount of data required depends on the nature of a given study, collecting social media data within a range of dates or from a particular geolocation, for instance, will limit the scope of a harvest while still supplying an adequate amount of data for analytics. As mentioned, sampling data offers a solution for reducing the size of a harvest but social media platforms are almost never forthcoming with their algorithms for how they select samples. Twitter's 1% streaming API, for instance, significantly reduces the number of tweets in a single request but it will not be possible to verify that the sample is random. Furthermore, filtering a harvest by keywords, or in some cases by hashtag, may also reduce the scope of a dataset, but there is no guarantee that such a filter will capture all relevant content. Similarly, the irregular use of keywords and hashtags may cause even a refined filter to capture irrelevant content. Filters may also fail to eliminate all obscenities, offensive content, or libellous content not wanted in permanent archives. These challenges are further complicated by the difficulty of detecting bots from real users, potentially polluting harvests with false data. The potential unreliability of using filters and samples to scope social media data poses a challenge for both research and heritage social media collections. Quality assurance could be a solution, but developing a method for checking the content of a social media harvest could be just as, or more, time consuming than the harvest.

Both restrictive selection policies (e.g. institutional accounts) and broad selection policies (e.g. topic or location filters) face the challenge of maintaining the meaning of the social media over time, which means ensuring that an archive contains enough metadata to provide meaningful context. In the case of Twitter, the Terms of Use include strict requirements about how to (re-)display individual tweets.[58] These requirements forbid the display of tweet content without the user name attached. If archived tweets do not contain relevant metadata and coherent context for individual tweets, users may be restricted as to how they can use the preserved data in the future.

[58] https://about.twitter.com/company/display-requirements

The secondary information attached to social media content, such as user ID, geolocation, user IDs for comments or shares, are especially critical for understanding archived content that is non-textual. Instagram photos or short Vine videos, for instance, may not have any linguistic information within the core content at all. However, much of the content published on social media derives from interaction with, or reference to, embedded URLs, images, audio clips, video, or other social media posts, such as posting a Vine to Facebook, or tweeting an Instagram photo, or sharing an article from an online news outlet. Many embedded URLs may also be shortened using services like TinyURL or Bitly. These shortened URLs may need to be converted to their original form before they can be preserved.

While maintaining the URLs that point to external content may offer short-term solutions, the only way to ensure that external content embedded in social media posts will be accessible over long periods of time is to preserve the external content simultaneously. The ARCOMEM project offers a strategy for combining the capture of social media through an API alongside the capture of linked web URLs through the integration of their harvesting tools (Risse, Peters *et al.*, 2014, p. 215). The UK Government social media archive uses an alternative solution of redirecting embedded central government URLs to the archived version of that website. These tactics preserve context as well as temporal cohesion, an issue facing the archiving of more traditional websites as well. Preserving social media means capturing enough content to provide meaning but also finding practical solutions to managing such large, diverse, and interlinked material.

While capturing *enough* relevant content poses one difficulty, ensuring that you do not capture *too much* also poses a challenge. Without a carefully planned and systematic strategy for de-duplication, an archive may end up with an abundance of redundant content that can cause increased difficulties for storage and search, further discussed in the next section. De-duplication can be managed by ensuring that relevant metadata and content remains linked to the tweet IDs for all data ingested into a database. Effective database management, therefore, will ensure de-duplication if data is maintained by tweet IDs or other persistent identifiers. In the case of archiving snapshots, de-duplication poses a far more difficult problem because it may not be possible to delete repeated data when snapshots are merged. Furthermore, Twitter's Developer Policy forbids the preservation of deleted tweets – thus an important part of a conversation may be missing from a harvest resulting in an incomplete record in the archive. Even if an institution or researcher harvests data from the API fast enough to capture content before it can be deleted (unlikely in most cases), the deleted content could compromise the institution's compliance with Twitter's policy.

5.6. Storage and search

Selection criteria are important not only for creating a meaningful and coherent collection of social media but also when considering long-term storage. Collections of social media come in many sizes and shapes, from spreadsheets containing tweet IDs to large corpora of social media harvested frequently over a long period of time. In 2013, the Twitter Archive at the Library of Congress was a hefty 80 terabytes of data containing more than 120 billion tweets (Library of Congress, 2013). The Social Data Science Lab at Cardiff University, Wales, has been harvesting data from Twitter's 1% streaming API daily for about three years, estimated at about two billion tweets and counting (see Case Studies for more details). The UK Government social media archive currently holds 65,000 tweets and 7,000 videos (Espley, Carpentier *et al.*, 2014, p. 46). These are all examples of collections that will grow fairly rapidly as the institutions collect and archive ongoing social media activity from these platforms. Even smaller archives which are only interested in discrete datasets around specific topics will face issues of growth over time as social media and web 2.0 continue to grow and cultural outputs migrate to new forms of media. Furthermore, as already mentioned, the terms and conditions of platforms may limit where and how data may be stored, prohibiting, for instance, storing data with a third party such as a cloud storage provider. Therefore not only the size but the legal framework of social media data influence how it can be stored.

While the size and relatively fast growth rate of social media data may demand large and scalable storage capacity, these characteristics also pose challenges to processing and indexing these data in order to render

them useable for current and future researchers. As demonstrated by the Twitter Archive at the Library of Congress, the difficulty of architecting a technical environment that allows users to query social media archives increases with the size of the data (Library of Congress, 2013). The Social Data Science Lab at Cardiff has also foreseen the approaching issue of indexing and processing as their archives grow (Burnap *et al.*, 2014). The team are currently working with HPC Wales to develop a technical solution that will ensure their researchers will be able to make use of the growing collection of social data, especially Twitter data (see Case Studies for more details). The ARCOMEM project, along with the Social Repository of Ireland and other social science-driven projects, sees the enrichment of Twitter data as an access mechanism for researchers, providing access points such as particular topics, keywords, or time intervals (Risse, Peters *et al.*, 2014, p. 209). Establishing the indexing and processing strategies for social media archives early in the process will be crucial to future usability of the data.

5.7. Long-term preservation

As with web archives, there is a pressing need to make a copy of social media for long-term preservation. In March 2015, web historian Peter Webster published a blog post on the rapid disappearance of web content: 'How fast does the web change and decay? Some evidence'.[59] In reference to the UK Web Archive, he reports that 'even for content that was archived only a year before, the proportion that is live and unchanged is less than 10%' (Webster, 2015). In particular, he quotes researchers SalahEldeen and Nelson (2012) who examined the lifespan of resources shared on social media and found that 'after the first year of publishing, nearly 11% of shared resources will be lost and after that we will continue to lose 0.02% per day' (quoted in Webster, 2015).

Social media data, like the resources it references, is also vulnerable to potential loss. In October 2014, social media users feared the disappearance of millions of photos uploaded to Twitpic. Twitter threatened to rescind API access to Twitpic unless the service shut down (D'Orazio, 2014). Only after a last-minute agreement did Twitter agree to take on the Twitpic domain and photo archive (*ibid.*). Commercial social media platforms, as discussed in Section 5.3 have a business model that values current data far more than historical data. Furthermore, social media platforms have not published internal preservation policies or plans. The array of sites rescued by the Archive Team provides some evidence of the track record of social media sites that have been closed, sometimes at short notice.[60] The reliance on an enthusiast-driven archiving initiative (albeit with support and direction from the Internet Archive) for securing this data suggests a gap in the priorities and collecting policies of traditional GLAM institutions (galleries, libraries, archives and museums). Given these conditions, users and policy makers have no reason to expect platforms to preserve user data for the long term.

At the time of writing, the practice of archiving social media by collecting institutions largely exists as an extension of web archiving, using similar tools to those used to harvest the traditional web. Most attempts at developing new methods to capture social media as machine-readable code have barely made it past the pilot stage. All of the case studies later in this report, for instance, represent work done in the last one to three years. Because of the novelty of archiving social media data, standards and best practice do not yet exist to benchmark the qualifications of a long-term strategy for preserving this content. While related standards provide some guidance, no one standard addresses the full range of activities needed to ensure the effective preservation of social media data and all relevant context.

On one hand, the preservation of social media data requires the capture and preservation of content data and metadata, for example a tweet and its user information in JSON (see Figures 1 and 2). On the other hand, the preservation of social media also requires the preservation of any embedded media or URLs. For instance, the URLs in a tweet may need to be converted to the original, full URL, and then linked to an

[59] http://webarchivehistorians.org/2015/03
[60] http://archiveteam.org/index.php?title=Category:Online_projects

archived version of the page. Alternatively, a YouTube video shared in a Facebook post may need to be harvested and preserved.

Social media data collected through APIs is often delivered as raw JSON or XML code. Twitter utilizes JSON, a 'lightweight, text-based, language-independent, data interchange format standardised by the Internet Engineering Task Force (IETF)' (Risse, Peters *et al.*, 2014, p. 217). JSON is an open standard derived from JavaScript. Alternatively, some social media APIs deliver data in XML, a non-proprietary, archival document encoding format based on ISO 8879-1986, now supported by W3C (2015).[61]

Twitter documents the structure of JSON formatting of tweets, but does not provide historical documentation of earlier versions of the tweet format.[62] Substantial changes were made in the move to version 1.1 of the Twitter API, and it seems likely that further changes will occur over time.[63] Not all social media platforms publish format documentation.

Preserving social media in JSON and XML does not provide a solution to ensure long-term access to contextual information that is merely referenced, such as websites or other media. For a platform like Twitter, which only allows posts of 140 characters, a missing URL or image may render a tweet or even an entire conversation completely meaningless. Furthermore, URLs may be shortened using a service like TinyURL or Bitly, adding a further layer of indirection – and another third-party service – that may be required to interpret social media content. A number of tools and methods have appeared that help to automate the conversion of shortened URLs to full URLs. The UK Government Twitter archive at The National Archives, for instance, uses a tool developed in partnership with the Internet Memory Foundation to detect shortened URLs and convert them to full URLs. Furthermore, the UK Government Twitter archive redirects all .gov.uk URLs to the archived version of the page. The archived version of the webpage is more secure than a live page and also provides a more chronologically accurate referent.

Overall, the capture and preservation of social media data requires adequate context. The experience of social media is diverse and the meaning of content heavily influenced by the means used to share it – the device of the user, the services offered by the platform, the appearance and features of the application, and all the related content and community around any interaction. The ability of future generations to understand archived social media in a meaningful way will require the preservation of user experiences – a task not always achievable through preserving content or metadata. Platforms frequently update the look and feel of their services and change basic functionality. For instance, Facebook changed its user profile-oriented interface to the 'Timeline' in late 2011, making it mandatory for all users in 2012,[64] and more recently Twitter changed its 'favourite' star to a 'like' heart.[65] Similarly, new platforms continue to appear as well as new devices and technologies for interacting with them.

Capturing data, metadata, and documentation may not provide enough context to convey user experiences with these platforms and technologies. One potential strategy for preserving user experience is to record user journeys or walkthroughs, much like the popular practice of screencasting in the gaming world – recording gameplay and uploading videos to YouTube, often with commentary.[66] The BBC Archive has made use of this strategy by 'creating screenshots and screencasts of BBC website user journeys and archiving selected high quality video and images which were created for the website'.[67] Rhizome's ArtBase has been preserving 'works of net art' in order to '[provide] a home for works that employ materials such as software,

[61] http://www.w3.org/standards/xml
[62] https://dev.twitter.com/overview/api/tweets
[63] https://blog.gnip.com/migrating-version1-1-twitter-api
[64] http://www.huffingtonpost.com/2012/01/24/facebook-timeline_n_1228800.html
[65] https://blog.twitter.com/2015/hearts-on-twitter
[66] http://www.theguardian.com/technology/2014/jan/02/lets-play-youtube-pewdiepie-one-direction
[67] http://www.bbc.co.uk/informationandarchives/archivenews/2014/archiving_bbc_online

code, websites, moving images, games, and browsers'.[68] The ArtBase focuses on the authenticity of user experience and develops approaches to digital preservation[69] that emphasize narrative over raw data.[70] The preservation of user experience could also potentially be archived through the preservation of hardware – of the laptops, phones, tablets, etc. used to create and consume this material, though this strategy would require the upkeep of old versions of operating systems and other hardware on increasingly ageing machines.

Whether through recordings or more complex approaches to capturing narratives of social media or web interactions, the preservation of user experiences will support the use and value of social media archives in future.

[68] http://rhizome.org/art/artbase
[69] https://rhizome.org/art/artbase
[70] http://blogs.loc.gov/digitalpreservation/2014/03/digital-culture-is-mass-culture-an-interview-with-digital-conservator-dragan-espenschied

6. Case Studies

6.1. Social Data Science Lab and the COSMOS Platform, Cardiff University

Among the increasing number of research initiatives that have emerged around new forms of web data, the COSMOS software platform initiative is one of the only projects focused specifically on social media data analytics and, perhaps more notably, on the development of new computational methodologies for using social media data in academic social science. The Social Data Science Lab at Cardiff University hosts and maintains the COSMOS platform, a freely accessible platform for not-for-profit use. The Lab's research poses a diverse range of questions but broadly addresses issues of criminality, tension, risk and wellbeing in online and offline communities.[71] Much of this research has been conducted using social media data in conjunction with more traditional forms of data, including a project using 15-year-old data from the UK Data Archive about the victims of crimes in order to classify Twitter data on more current crimes. These studies, alongside increased experience working within the restrictions of the Twitter Terms and Conditions, have enabled researchers at the Social Data Science Lab to construct and test new methods and procedures for performing academic social science on social media data. In order to support this new framework, the Lab has developed a strategy for capturing and storing this data to meet the current and future needs of social science research. Though currently restricted from sharing this data, the tools and infrastructure developed for the COSMOS platform provide useful insight into the data needs of researchers, such as the types of data, metadata, and formats required to perform analytics.

As part of the overall framework for analysing social data, the Social Data Science Lab has developed the COSMOS open source software for non-commercial use; this will help facilitate large-scale social media data analytics, the largest social data source being Twitter. Using APIs, COSMOS ingests raw Twitter data and metadata into a NoSQL database. The original data remains in its raw format and is never altered, in order to ensure the integrity and authenticity of the data over time. In order to index and query the raw JSON data, COSMOS has added a database layer populated with additional attributes derived from the JSON metadata. This added layer makes it easier to analyse the data in different ways. COSMOS has been pulling 1% of the streaming API daily for about three years, estimating about two billion tweets. The team has also purchased data from Twitter surrounding major events. In addition to these tweets, over a period of about 18 months COSMOS collected geo-located data from across the UK, estimating 500,000 data points daily. All of these data are stored in databases held on local servers.

The Social Data Science Lab, like all organizations and individuals who access Twitter data directly from an API, abides strictly by the social media platform terms and conditions and developer policies. This means that the Lab cannot share any data from its Twitter archive nor can they hold any of their archive in cloud storage. As an observatory dedicated to developing the use of social data for social science, however, the Lab has entered communications with Twitter in hopes of negotiating greater access to the COSMOS platform's data more regularly. As part of this initiative, the Lab aims to collaborate with other organizations to increase access to social media data and enable sharing across organizations. In addition, they have begun working with HPC Wales, a national supercomputing service provider, to develop stronger technology for processing and querying social media data as it inevitably grows. This progress in using social media in academic social science establishes a precedent for new academic projects as well as for collecting institutions interested in building repositories to accommodate social media data. Advances in the computational analysis of social media data based on refined methodologies and standard practices could provide useful prototypes for the long-term preservation of this type of digital content. The structure of databases, for instance, and the enrichment of data with added attributes in order to maintain necessary context for understanding the content of social media posts over long periods of time are also effective approaches to long-term preservation.

For more information about the Social Data Science Lab, please visit: http://socialdatalab.net

[71] http://socialdatalab.net/publications

6.2. Social Repository of Ireland feasibility study (Insight @ NUI Galway & The Digital Repository of Ireland)

The Social Repository of Ireland was a one-year feasibility study to explore the technical and archival challenges in creating a richly annotated archive of social media responses to major events and topics of interest in Ireland, and to preserve this archive in a trusted digital repository – The Digital Repository of Ireland (DRI). In addition to archiving and preservation, the study developed practices for semi-automating the processes of annotation and archiving, enriching the data with contextual curation from archives, and assessing legal and ethical issues that arise in the preservation of social media.

As a pilot project, the Social Repository exclusively collects Twitter data, due to the relative ease of accessing this through Twitter's APIs. The Insight Centre for Data Analytics at NUI Galway has created a suite of tools that filter data from the APIs by relevance to Ireland, using hashtags and keywords to restrict incoming tweets to those within the national scope of the project. The Social Repository project implements these tools and methods to organize and annotate the social media in its collection in a way that will make it meaningful to current and future users. These methods include the automation of the collection process by detecting topic-specific content, such as Irish public figures, geographical locations, and prominent institutions. The suite of tools also applies a classification scheme that allows users to find content by topic, such as sport, politics, or major event, using metadata provided by Twitter. The Social Repository holds, for instance, public tweets surrounding the historic Constitutional Referendum to recognize same-sex marriage held in Ireland in May 2015. After collecting these tweets, Insight transfers the enriched datasets to DRI who preserve the data in order to ensure long-term access for journalists and researchers. The project treats social media as a significant aspect of a nation's historical and cultural record – an aspect worth preserving alongside more traditional records, such as newspaper archives.

The feasibility study focused on the particular constraints and potential solutions to extending and maintaining such a repository in the long term. The Social Repository project has demonstrated, for instance, some strategic methods of providing access to the Twitter data collected by Insight, given the strict terms of Twitter's Developer Agreement & Policy. An interview with Clare Lanigan, the Social Repository's Research Assistant at the DRI, revealed some of the issues with making archived Twitter data accessible. To 'follow the rules', for example, the Social Repository would hold raw data that can only be accessed by authenticated academic researchers with a verifiable institutional login. Once researchers have accessed the data, they will be able to filter data based on topic with a user-friendly interface and download a spreadsheet containing a selected dataset in a program such as Excel. The access system limits the number of downloads per user, further securing the archived data. These access policies, alongside a detailed discussion of Twitter's API regulations in a blog post about the Social Repository project, indicate an exhaustive effort to comply with the platform's rules while finding innovative ways of providing Ireland's researchers and journalists with valuable source data. As Lanigan concludes in her post about Twitter's historical API regulations: 'no archive of social media can afford to neglect looking for solutions to the problems of collecting, preserving and making this data available'.

The Social Repository of Ireland feasibility study provides a model for effective collaboration between data science and digital archives. Insight at NUI contribute a mature knowledge of the tools and technical mechanisms for capturing social data for analytics, and specifically, for supporting data journalism. The DRI contributes a sophisticated knowledge of archiving social media content and maintaining long-term access to Ireland's digital cultural heritage. The suite of tools developed by Insight harvests social data with archiving in mind, while DRI has developed methods and processes for receiving data deposited by data scientists. In addition to the data, DRI will in the future host the tools developed by Insight, ensuring the sustainability of the project outputs. This model provides a template and a lesson in cross-sector co-operation for any institution, big or small, interested in studying or archiving social media.

6.3. GESIS Leibniz Institute for the Social Sciences

Since 1986, the GESIS Institute for the Social Sciences, an independent institution in Germany for the curation of deposited data, has provided access to social science data to support the reproducibility of research.[72] It uses international standards for research data management and for data protection and privacy. The Institute performed a pilot study into archiving social media through collecting and archiving social media data from Facebook and Twitter around the 2013 German Bundestag (parliamentary) elections (Kaczmirek, Mayr *et al.*, 2014). Using the Twitter APIs, GESIS collected relevant tweets, filtered by hashtag and date range. For Facebook data, GESIS collaborated with the Copenhagen Business School's Computational Social Science Laboratory (CSSL), using their Social Data Analytics Tool (SODATO) to harvest and process graph data and social text data (*ibid.*, pp. 2, 10).

Through this pilot study, GESIS took on common challenges to the re-use of social media research data, including the lack of standards and documentation but also the difficulty that 'data from social media platforms are often unavailable for secondary use' (*ibid.,* p. 4). By collecting and archiving Facebook and Twitter data around the 2013 Bundestag elections, GESIS created a social media dataset that could serve as source data for analytics. The specialists at GESIS were then able to explore the practical options for archiving and disseminating this data.

In the pilot study, GESIS developed two alternative approaches to the archiving and dissemination (or access method) of the Bundestag dataset, relying on its established expertise in archiving sensitive social science survey data as well as on the relevant guidelines provided in the Twitter Developer Agreement & Policy. Based on experience of archiving sensitive survey data, GESIS is exploring the possibility of allowing access to the dataset through a secure data centre, either on site or through authenticated remote login (*ibid*., p. 18). This method would allow authenticated researchers to work with data that has not been fully anonymized (an incredibly time-consuming task with such large, inter-linked data). Alternatively, for Twitter data, GESIS now provide access to the tweet IDs for a given dataset (Kaczmirek and Mayr, 2015). Twitter's Developer Agreement & Policy specifically articulates this method of sharing Twitter data, making it a safe option for institutions taking great pains to ensure compliance with the platform's terms and conditions (Kaczmirek, Mayr *et al.*, 2014, p. 18). Both methods may pose difficulties for authenticity. The secure data centre poses fewer difficulties, but limits the validation of data to a few authenticated researchers. Tweet IDs pose more difficulties for authenticity, as the process of re-creating a dataset based on tweet IDs cannot account for tweets deleted from the original dataset. Tweet IDs also cannot guarantee the exact replication of methods for collecting and cleaning (or curating) used in the original dataset (Summers, 2015). Furthermore, the preservation of tweet IDs relies on Twitter, or another source, maintaining this data for the long-term.

These strategies for archiving social media content (used by both GESIS and other institutions), provide excellent guidance for institutions that support social media researchers. The methods are most useful for Twitter data but can also be adapted to a range of social media platforms and research disciplines. The approaches taken by GESIS can also, however, provide useful techniques for institutions that do not currently have an active user community for archived social media content. Heritage sector institutions, for instance, could rely on this strategy by keeping an additional metadata object attached to the social media content itself. This related metadata object could facilitate easier sharing in future, as long as the data is still available through the relevant platform or another archive (such as that, potentially, at the Library of Congress). The Legal Deposit Libraries in the UK and Ireland already provide a similar service as that of a Secure Data Centre for the UK Web Archives, an extension of which may provide solutions for access to heritage social media collections (British Library, 2013). Though the legal constraints around the sharing of social media data create significant impediments to maintaining long-term access, the development of uniform methods of documentation offer a practical solution that can be implemented now.

[72] http://www.gesis.org/en/institute

6.4. The National Archives' UK Government Social Media Archive

The National Archives in partnership with the Internet Memory Foundation (IMF) launched a pilot project in 2011 to capture official central government communications on the most heavily used social media platforms – Twitter and YouTube (Storrar, 2014).[73] Within the framework of their existing UK Government Web Archive (UKGWA) programme, which began in 2003, the UKGWA team and IMF developed an archiving approach based on the harvesting of data directly from platform APIs (Espley, Carpentier *et al.*, 2014). The objective of the pilot was to find a solution to archiving social media content that could overcome the problems of traditional methods. Before the implementation of the API-based solution, 'those websites containing content published on [social media] would either not be fully captured during the archiving process or would be captured but inaccessible to the user of the archive as a result of their file size' (*ibid.*, p. 33). The pilot study developed a selection policy and technical approach to allow The National Archives and IMF to cater for the acquisition, curation, and redeployment of these records within relevant legal and policy frameworks and to meet the needs of the UKGWA's user community. This catered approach supplied the team at TNA with a means to select only that content and metadata published by official central government accounts in order to fulfil the obligation created by the Public Records Act 1958 to preserve all government records regardless of format and to simultaneously avoid the risk of infringing copyright outside of Crown Copyright (*ibid.*, pp. 33–38).

In order to implement this API-based solution, the UKGWA and IMF collaborated to develop appropriate technical solutions, including tools and players, to make the process of archiving social media as automated as possible. The pilot project created tools to ensure that only the relevant data and metadata are harvested from the Twitter API, and a set of processes that also has the capability to separate link shortening services that resolve to UK government URLs within tweets (such as .gov.uk), from URLs to external, non-government, sites. This tool allows TNA to direct users to the archived version of relevant URLs within the UKGWA and also provides a 410 error message to refer users to the live web for external sites. To replay tweets, the UKGWA website displays the IMF Twitter Archive Reader that 'echoes the Twitter.com interface' and allows users to experience the tweets in a comparable way to the originals (*ibid.*, p. 42). In addition, users can download the original JSON and XML data underlying these tweets for future processing and analysis. YouTube videos are played back on the UKGWA website through JW Player, an open source player, where relevant metadata is also displayed, including tags and descriptions of videos. Acknowledging the challenges of defining authenticity of archived web content and establishing the trust of users, the UKGWA makes all original metadata visible to users and openly discloses their selection and archiving process (*ibid.*, p. 42).

This strategy for archiving social media and making it accessible as machine-readable data provides a useful model for institutions with a legislative or regulatory obligation to maintain long-term access to their official records, particularly if they have a requirement to make these records open to the public. The use of APIs and web-based players that imitate (but do not replicate) the experience of interacting with the original content may have a broader application as different types of institutions develop policies and strategies for complying with commercial platform terms and conditions. As the social media collections in the UKGWA grow and develop, the collection may continue to provide guidance on the development of API-based social media archives. The team seeks to develop the search functionality and indexing of the archived tweets and YouTube content and explore possible integrations with other web archives, such as the UK Web Archive and Internet Archive, to improve the user journey between UKGWA and other archived web content that falls within the remit of those institutions. These initiatives should improve users' experience of the social media archive in the future. Both improved search and inter-linking could enhance the meaning and significance of archived social media by providing a wider cultural and social context. Furthermore, this collaborative model could support shared knowledge of best practice for preserving social media and increased availability of social media archives for users.

[73] http://www.nationalarchives.gov.uk/webarchive/twitter.htm and http://www.nationalarchives.gov.uk/webarchive/videos.htm

7. Conclusions and Recommendations

Social media data and analytics is a new area of study in the social sciences and other academic disciplines; so new that this field has only just begun to develop relevant methodologies and standard practices. In the heritage and government sectors, also, social media is only beginning to take shape as a cultural artefact and official record. The underlying technologies that underpin social media platforms change continuously, with new platforms and applications appearing just as quickly. Platforms are updating their terms and conditions so frequently that, in many cases, tools and applications developed to interact with this data are restricted or forced to shut down after only a few years. As a result, strategies for preserving social media are largely adaptive, relying on established practices for analogous types of digital content, such as the traditional web, research data from other sources, licensed content, and records containing personal or sensitive information. The 'social', or dynamic, web, will continue to grow and mature –likely evolving into new forms of technology and content – and will require flexible, co-ordinated archiving strategies capable of operating inside a complex legal and ethical framework. Currently, only a relatively small number of people possess the technical skills to develop, manipulate, and access social media data through APIs. The requirement for this skillset will have an impact on new developing policies and practices around social media archiving.

As research and collecting institutions develop policies and strategies for archiving social media, the evaluation of user needs and expectations will help guide decisions about what to preserve, how much, and in what form. For research institutions that support data analytics and academic research using social media data, the involvement of researchers and scientists in early stages of development will provide invaluable information about how to select and curate social media data. This early consultation with users will also help establish stronger relationships between researchers and the archives that support them. Some social media archives have already conducted research into the needs of their users, in both the research and the archive sectors. North Carolina State University Libraries, for example, have published the results of a survey of archival researchers into the use of social media in academic studies.[74] These types of research may also support outreach and communications to engage researchers and make them aware of the social media data resources available for use. This outreach could also contribute more widely to data literacy and greater public awareness of what happens to user data once it is published on a platform.

Transparency and openness will also be important, not only for legal and ethical requirements but for practical reasons as well. Developing policies and procedures for capturing, organizing, and analysing social media data publically available will make it possible for social media users to find information about how their data might be used beyond the public interfaces of platforms. However, it will also provide critical metadata for re-creating datasets in instances where data collected by researchers cannot be deposited in the archive. Though re-creating a dataset, even based on detailed documentation, does not guarantee an identical dataset, it is currently the only practical substitute. Furthermore, transparency in the process of capture, organization (or 'cleaning'), and analysis of datasets provides important provenance for archives that will maintain the data over the long term. This type of information also supports rights management and ensures that the rights attached to a particular dataset (for instance, the terms of a developer agreement when the data was originally captured) will be available to future researchers. In addition, any available documentation for the tools and programmes used to interact with social media data should be preserved alongside a dataset. This information will help ensure future researchers have the information they need to understand how a dataset was created and how users would have experienced social media at the time s historical dataset was generated. Sharing research methods for working with social media data as well as sharing the policies and procedures for archiving that data and its accompanying metadata reflects trends in the research sector that encourage, or require, open data but may also support closer collaboration between researchers and collecting institutions as well as closer collaboration among collecting institutions.

[74] North Carolina State University, Social Media Archives Toolkit: https://www.lib.ncsu.edu/social-media-archives-toolkit/surveys/researcher-survey

In the effort to maintain long-term access to social media, alongside other forms of big data, close collaboration will help surmount some of the greatest challenges, such as the volume and variety of social media research datasets. Collaboration among similar institutions with overlapping remits, such as universities or national heritage libraries, could involve sharing costs and widening access to valuable datasets. This collaboration might involve joint projects or developing a shared technical infrastructure. A more substantial co-ordinated effort would be to integrate social media archives with traditional web archives. This integration of archived web resources would make it possible to refer URLs in social media to the archived version of a webpage. Linking social media data to its larger context on the web would provide more authentic records for current and future researchers. Overall, collaboration across multiple institutions will facilitate the sharing of knowledge and improve social media archiving. To consolidate and support collaboration across multiple institutions, the development of one or a few centralized infrastructures could help improve the standardization of practices that can facilitate better preservation.

When considering the big picture, however, the preservation of social media may best be undertaken by a large, centralized provider, or a few large centralized providers, rather than linking smaller datasets or collections from many different institutions. For the purposes of data analytics, archives will need to preserve terabytes or petabytes of data. While there is also great value in computational or qualitative analyses that look at smaller datasets, or even individual posts, both large- and small-scale data research is served by large, aggregate collections of data. There are a number of other advantages to a large, centralized infrastructure for social media data besides the necessity for large-scale analytics. In 2013–14, researchers Weller and Kinder-Kurlanda performed a widespread survey of social media researchers where participants 'critically discussed the possibility of one or more centralized research institutions who could serve as providers of social media research data' (2015, p. 34). Some researchers were in favour of this type of centralized infrastructure because it would provide benchmarks for data quality as well as reduce data costs for individual researchers. Some sceptics, recognizing the disadvantages of consolidating all social media archives into a single institution, 'argued that a central social media data provider for researchers would … need mechanisms to ensure reliability, comparison and interchangeability of the data' (Weller and Kinder-Kurlanda, 2015, p. 34). In order to meet these provisos, it makes sense to develop an infrastructure either within, or governed by, a national institution that already acts as a source of guidance to best practice in the area of data management and preservation.

Centralizing social media data does not mean the obligatory deposit of all social media archives with a single organization, rather it provides a source for researchers and collecting institutions to build or develop their own social media archives needed for their own local collections. A centralized infrastructure could liaise or negotiate with social media platforms in order to align more closely research standards and requirements, legal deposit regulations, and, where possible, the terms and conditions of social media platforms. A centralized infrastructure would also be able to facilitate the harmonization of social media collecting policy and standards across multiple institutions. While local social media archives, and the personal collections kept by authors and researchers, contain significant value, in order to establish significant progress in the capture and preservation of social media, infrastructure and standardization must be developed at a higher level. This development will require collaboration both among collecting institutions, but also across sectors that have a stake in long-term access to social media.

8. Glossary

*Definitions with an asterisk derive from the OECD Ethics Glossary, and may originate from other sources (those sources are cited in the relevant definitions).

Access token: A single permission needed to make requests from APIs, provided by platforms in limited quantities in order to prevent developers from requesting too much data or from requesting data too frequently.

Analytics (data analytics): The computational analysis of large sets of data, from one or many sources, to discover patterns or trends, such as about an event, phenomenon, or demographic. Discerns information within the data not visible without the hardware and software technology to manage, store, process, analyse, and synthesize very large sets of data.

Anonymization*: A process of ensuring that the risk of somebody being identified in the data is negligible. This invariably involves more than simple de-identification, but also requires that data be altered or masked in some way in order to prevent statistical linkage.

Application Programming Interface (API): A set of routines, protocols, and tools designed for developers to build applications on top of the underlying building blocks of an original piece of software, such as Twitter or Facebook, allowing the underlying software to hide particular implementations while still sharing the information needed to create new applications. In the context of this report, an API provides an interface for communicating with the back end of a social media system and enabling such functions as querying or requesting copies of social media data.

Consent*: Informed consent entails giving prospective participants sufficient information about the research and ensuring that there is no explicit or implicit coercion so that they can make an informed and free decision on their possible involvement. Information should be provided in a form that is comprehensible and accessible, typically in written form (or in a form that participants can access after the end of the research interaction), and time should be allowed for the participants to consider their choices and to discuss their decision with others if appropriate. The consent forms should be signed off by the research participants to indicate consent. (Source: ESRC Framework for Research Ethics)

CSV: A comma-separated values (CSV) file stores tabular data (numbers and text) in plain text. Each line of the file is a data record consisting of one or more fields, separated by commas. There is no official standard for the CSV file format, but RFC 4180 provides a de facto standard for many aspects. CSV is supported by a very large number of tools, from spreadsheets such as Excel, OpenOffice and Google Docs to complex databases to almost all programming languages. (Sources: Open Knowledge http://data.okfn.org/doc/csv and Wikipedia https://en.wikipedia.org/wiki/Comma-separated_values)

Data cleaning (or cleansing)*: Process to check data for adherence to standards, internal consistency, referential integrity, valid domain, and to replace/repair incorrect data with correct data. To 'clean' a data file is to check for wild codes and inconsistent responses; to verify that the file has the correct and expected number of records, cases, and cards or records per case; and to correct errors found. (Source: ICPSR Glossary)

Data owner*: A legal entity which has the right to give permission for rights (intellectual property) to be used by others. A 'data owner' could be an individual or a corporation.

Data science: A cross-disciplinary research approach that uses large amounts of data for analysis; 'data science' is used by government, journalism, business, academic social science, computer science, and even by the humanities.

Developer: Also referred to as a software engineer; a person who writes computer programmes and has the required skills and knowledge to build or use tools to access platform APIs and request data, as well as to organize and query that data.

Developer policy: see Terms and Conditions.

Disclosure (accidental disclosure)*: Disclosure relates to the inappropriate attribution of information to a data subject, i.e. an individual person or organization represented in a set of data. Disclosure has two components: identification and attribution. (Source: OECD Expert Group for International Collaboration on Microdata Access: Final Report)

JSON: JavaScript Object Notation. A data interchange format standardized by the Internet Engineering Task Force (IETF). JSON is a popular language format because it is compact, text-based, and language-independent. It is an open standard derived from JavaScript. (Source: Risse, T *et al.,* 2014.)

NoSQL or Non-Relational Database: A way of organizing a database that does not use tabular relations to organize individual items of data. Though there are multiple types of NoSQL database designs, the design most useful for storing social media data is document-oriented databases. Document-oriented databases allow developers to store and query documents encoded in standard encodings or formats such as XML or JSON. MongoDB is a popular type of NoSQL database because it provides index support and straightforward queries. (Source: https://en.wikipedia.org/wiki/NoSQL#Document_store and Kumar, S *et al.*, 2014)

Open data*: Data (datasets) that are: 1) accessible to anyone and everyone, ideally via the Internet; 2) in a digital machine-readable format that allows interoperation with other data; 3) available at reproduction cost or less; and 4) free from restrictions on use and re-use. (Source: OECD Expert Group for International Collaboration on Microdata Access)

Persistent identifier: A long-lasting reference to a digital resource. Typically it has two components: a unique identifier; and a service that locates the resource over time even when its location changes. (Source: Digital Preservation Handbook, http://www.dpconline.org/advice/preservationhandbook/technical-solutions-and-tools/persistent-identifiers)

Personal data*: Any information relating to an identified or identifiable individual. (Source: OECD Privacy Framework)

Platform: The hardware and software, including databases and web-based programming, which underlie applications such as social media services.

Privacy*: Someone's right to keep their personal matters and relationships secret, involving an obligation of the holder of information to the subject of the information to do so. (Source: UNECE, Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes, 2009)

Real-time: see 'streaming'.

Request: To initiate the transfer of data from an API, usually requires access tokens.

Streaming or real-time API: As opposed to a REST API, which requires querying a platform's stores of data, a streaming API connects using a persistent HTTP connection. Provides access to user-generated content being created in the present time, so that data 'streams' from the backend of the platform into the developer's storage medium. Twitter describes its streaming API as 'streams of the public data flowing through Twitter. Suitable for following specific users or topics, and data mining'. (Source: https://dev.twitter.com/streaming/overview)

Terms and conditions (platform terms and conditions): A set of restrictions under which the user of a service agrees to operate. They typically include: 1) the use and re-use of content; 2) underlying code; and 3) identifying artwork or branding from a social media platform, often including terms of service, user agreements, and developer policies. Developer policies and agreements often stipulate how the data requested from an API can be accessed, used, and displayed.

Web 2.0: A term coined to refer to interactive, social and collaborative web technologies, resulting in a more distinctive and more modern World Wide Web.

XML (eXtensible Mark-up Language): A widely used application-independent mark-up language for encoding data and metadata.

9. Further Reading

9.1. Case Studies

Social Data Science Lab and COSMOS Platform Case Study
Burnap, P, Rana, O, Williams, M, Housley, W, *et al.* (2014), 'COSMOS: Towards an Integrated and scalable service for analysing social media on demand', *International Journal of Parallel, Emergent and Distributed Systems*, 30:2, 80–100, DOI: 10.1080/17445760.2014.902057

INSIGHT@NUI Galway's Social Repository of Ireland at the Digital Repository Ireland (DRI)
The Irish Times (5 November 2014), 'All the data that's fit to print – and archive for posterity', http://www.irishtimes.com/news/science/all-the-data-that-s-fit-to-print-and-archive-for-posterity-1.1965955

GESIS Leibniz Institute for the Social Sciences
Kaczmirek, L, Mayr, P, Vatrapu, R, *et al.* (31 March 2014), 'Social Media Monitoring of the Campaigns for the 2013 German Bundestag Elections on Facebook and Twitter', GESIS Working Papers http://arxiv.org/abs/1312.4476v2

The National Archives' UK Government Social Media Archive
Espley, S, Carpentier, F, Pop, R, Medjkoune, L (August 2014), 'Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content', *Alexandria: The Journal of National and International Library and Information Issues*, 25:1–2, pp. 31–50(20). DOI: http://dx.doi.org/10.7227/ALX.0019

9.2. Tools and Resources

ARCOMEM (open source)
About: http://sourceforge.net/projects/arcomem

COSMOS (open source)
About: http://www.cs.cf.ac.uk/cosmos
To download: http://www.cs.cf.ac.uk/cosmos/download-cosmos

NCSU Social Media Archives Toolkit (resource)
https://www.lib.ncsu.edu/social-media-archives-toolkit

NCSU Tools (open source)
Lentil: https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/lentil-user-guide
Social Media Combine: https://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/combine-user-documentation

Overview of Research Tools for Twitter and other platforms
'Using Twitter as a data source: An overview of current social media research tools' by Wasim Ahmed
http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview

Social Feed Manager (open source)
http://social-feed-manager.readthedocs.org/en/latest/

SODATO (proprietary)
About: http://link.springer.com/chapter/10.1007/978-3-319-06701-8_27
To Create an Account: http://cssl.cbs.dk/software/sodato

TWARC (open source tool)
To download: https://github.com/edsu/twarc

10. References

Ahmed, W 10 July 2015, 'Using Twitter as a data source: An overview of current social media research tools', The Impact Blog, London School of Economics and Political Science http://blogs.lse.ac.uk/impactofsocialsciences/2015/07/10/social-media-research-tools-overview

Archive Team 2009, 'GeoCities', ArchiveTeam.org http://www.archiveteam.org/index.php?title=GeoCities

Bandziulis, L 15 July 2014, 'How to Download and Archive Your Social Media Memories', WIRED.com http://www.wired.com/2014/07/archive-social-networks

Banks, M 2008, *On the Way to the Web: The secret history of the Internet and its founders*, Berkeley, CA: Apress. (E-book)

British Library October 2013, 'Accessing Web Archives' http://www.bl.uk/catalogues/search/pdf/accessing_web_archives.pdf)

Brustein, J 1 October 2014, 'Twitter Gives MIT $10 Million to Study the Social Impact of Tech', Bloomberg Business, http://www.bloomberg.com/bw/articles/2014-10-01/twitter-gives-mit-10-million-to-study-the-social-impact-of-tech

Burnap, P, Rana, O, Williams, M, Housley, W, *et al.* 2014, 'COSMOS: Towards an integrated and scalable service for analysing social media on demand', *International Journal of Parallel, Emergent and Distributed Systems*, 30:2, 80–100, DOI: 10.1080/17445760.2014.902057

Cate, F and Mayer-Schönberger, V 2012 'Notice and Consent in a World of Big Data', Microsoft Global Privacy Summit Summary Report and Outcomes http://download.microsoft.com/download/9/8/F/98FE20D2-FAE7-43C7-B569-C363F45C8B24/Microsoft%20Global%20Privacy%20Summit%20Report.pdf

Digital Curation Centre, 'Funders' data policies', DCC website, http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

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

D'Orazio, D 25 October 2014, 'Twitpic saved by Twitter just hours before planned shut down', *The Verge* http://www.theverge.com/2014/10/25/7070585/twitpic-saved-by-twitter-just-hours-before-planned-shut-down

Espley, S, Carpentier, F, Pop, R, Medjkoune, L August 2014, 'Collect, Preserve, Access: Applying the Governing Principles of the National Archives UK Government Web Archive to Social Media Content', *Alexandria: The Journal of National and International Library and Information Issues*, 25:1–2, pp. 31–50(20). DOI: http://dx.doi.org/10.7227/ALX.0019

Executive Office of the President May 2014, 'Big Data and Privacy: A Technological Perspective', Report to the President, President's Council of Advisors on Science and Technology https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

Foursquare (last updated 5 November 2014), 'API Platform and Data Use Policy' https://foursquare.com/legal/api/platformpolicy

Gates, C 4 June 2015, 'Eulogy for Politwoops', Sunlight Foundation blog, http://sunlightfoundation.com/blog/2015/06/04/eulogy-for-politwoops

Gillis, M 1 October 2014, 'Investing in MIT's new Laboratory for Social Machines', Twitter blog, https://blog.twitter.com/2014/investing-in-mit-s-new-laboratory-for-social-machines

Google+ (last updated 26 February 2013), 'Platform Terms of Service' https://developers.google.com/+/web/terms

Helmond, A 23 September 2015, 'The Web as Platform: Data Flows in Social Media', PhD dissertation, University of Amsterdam http://www.annehelmond.nl/wordpress/wp-content/uploads//2015/08/Helmond_WebAsPlatform.pdf

Housley, W and Williams, M, *et al.* (Eds) 2013, 'Computational Social Science: Research Strategies, Design and Methods', *International Journal of Social Research Methodology*, Special issue, 16, 2.

Hockx-Yu, H 2014, 'Archiving Social Media in the Context of Non-print Legal Deposit', *IFLA WLIC Libraries, Citizens, Societies: Confluence for Knowledge in Lyon* http://library.ifla.org/999/1/107-hockxyu-en.pdf

Kaczmirek, L, Mayr, P, Vatrapu, R, *et al.* 31 March 2014, 'Social Media Monitoring of the Campaigns for the 2013 German Bundestag Elections on Facebook and Twitter', *GESIS Working Papers* http://arxiv.org/abs/1312.4476v2

Kaczmirek, L, and Mayr, P 2015, 'German Bundestag Elections 2013: Twitter usage by electoral candidates.' *GESIS Data Archive*, Cologne, DOI: 10.4232/1.12319

Koops, B 2011, 'Forgetting Footprints, Shunning Shadows. A Critical Analysis of the "Right To Be Forgotten" In Big Data Practice.' *SCRIPTed*, 8:3, 229-256, DOI: 10.2966/scrip. 080311.229 http://script-ed.org/wp-content/uploads/2011/12/koops.pdf

Kumar, S, Morstatter, F, Liu H 2014, 'Twitter Data Analytics', DOI 10.1007/978-1-4614-9372-3

Lanigan, C 29 May 2015, 'Archiving Tweets: Reckoning with Twitter's Policy', *Insight News Lab* http://newslab.insight-centre.org/tweetarchivingchallenges

Library of Congress April 2010, 'Twitter Donates Entire Tweet Archive to Library of Congress', News Releases http://www.loc.gov/today/pr/2010/10-081.html

Library of Congress January 2013, 'Update on the Twitter Archive at the Library of Congress', White Paper http://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf

Mantelero, A 2013, 'The EU Proposal for a General Data Protection Regulation and the roots of the "right to be forgotten."' *Computer Law & Security Review*, 29:3, 229–235. http://dx.doi.org/10.1016/j.clsr.2013.03.010

McLemee, S 3 June 2015, 'The Archive Is Closed', InsideHigherEd.com https://www.insidehighered.com/views/2015/06/03/article-difficulties-social-media-research
Messerschmidt, J 15 April 2014, 'Twitter welcomes Gnip to the Flock', Twitter blog https://blog.twitter.com/2014/twitter-welcomes-gnip-to-the-flock

MIT News 1 October 2014, 'MIT launches Laboratory for Social Machines with major Twitter investment', http://news.mit.edu/2014/twitter-funds-mit-media-lab-program-1001

North Carolina State Universities (NCSU) Libraries 2014–15, 'Social Media Archives Toolkit', https://www.lib.ncsu.edu/social-media-archives-toolkit

Organisation for Economic Co-operation and Development (OECD) February 2013, 'New Data for Understanding the Human Condition', OECD Global Science Forum Report http://www.oecd.org/sti/sci-tech/new-data-for-understanding-the-human-condition.pdf

Pawlik, A, Petrie, M, Segal, J, Sharp, H 2015 'Crowdsourcing Scientific Software Documentation: A Case Study of the NumPy Documentation Project', *Computing in Science & Engineering*, 17:1, pp.28–36.

Pennock, M 2013, *Web-Archiving*. *DPC Technology Watch Report* 13-01 http://dx.doi.org/10.7207/twr13-01

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.

Risse, T, Peters, W, Senellart, P, and Maynard, D 2014, 'Documenting Contemporary Society by Preserving Relevant Information from Twitter', in Weller, K, *et al.* (Eds), *Twitter and Society*, NYC, NY: Peter Lang Publishing.

SalahEldeen, H and Nelson, M 2012, 'Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?', http://arxiv.org/abs/1209.3026

Schroeder, R December 2014, 'Big Data and the brave new world of social media research', *Big Data & Society*, DOI: 10.1177/2053951714563194
http://bds.sagepub.com/content/1/2/2053951714563194

Scola, N 11 July 2015 'Library of Congress' Twitter Archive is a Huge #FAIL', Politico.com, http://www.politico.com/story/2015/07/library-of-congress-twitter-archive-119698.html

Sloan L, Morgan J, Burnap P, Williams M 2015, Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data. *PLoS ONE* 10:3: e0115545, DOI: 10.1371/journal.pone.0115545

Summers, E 14 April 2015, 'Tweets and Deletes: Silences in the Social Media Archive', *Medium* blog post, https://medium.com/on-archivy/tweets-and-deletes-727ed74f84ed

Storrar, T 8 May 2014, 'Archiving social media', TNA Blog http://blog.nationalarchives.gov.uk/blog/archiving-social-media

Twitter, Developer Agreement & Policy, last updated 18 May 2015
https://dev.twitter.com/overview/terms/agreement-and-policy

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-for-social-and-economic-research/

US Government Accountability Office 31 March 2015, 'Library of Congress: Strong Leadership Needed to Address Serious Information Technology Management Weaknesses', Report to Congressional Committees http://www.gao.gov/assets/670/669367.pdf

W3C 2015, 'XML Technology', Standards http://www.w3.org/standards/xml

Webster, P 20 March 2015, 'How fast does the web change and decay? Some evidence', W*eb Archives for Historians* blog http://webarchivehistorians.org/2015/03

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

Zimmer, M 6 July 2015, 'The Twitter Archive at the Library of Congress: Challenges for information practice and information policy', *First Monday*, 20:7, DOI: http://dx.doi.org/10.5210/fm.v20i7.5619