Preserving Social Media: the Problem of Access

1st Annual Conference on Digital Preservation for the Arts, Social Sciences and Humanities (DPASSH 2015)

25-26 June, 2015

Croke Park Conference Centre, Dublin, Ireland

Sara Day Thomson
Digital Preservation Coalition
11 University Gardens
Glasgow G12 8QQ
sara.thomson@dpconline.org

Dr. William Kilbride
Digital Preservation Coalition
11 University Gardens
Glasgow G12 8QQ
william.kilbride@dpconline.org

ABSTRACT

As the applications and services made possible through Web 2.0 continue to proliferate and influence the way individuals exchange information, the landscape of social science research, as well as research in the humanities and the arts, has the potential to change dramatically and to be enriched by a wealth of new, user-generated data. In response to this phenomenon, the UK Data Service have commissioned the Digital Preservation Coalition to undertake a 12-month study into the preservation of social media as part of the 'Big Data Network' programme funded by the Economic and Social Research Council (ESRC). The larger study focuses on the potential uses and accompanying challenges of data generated by social networking applications.

This paper, 'Preserving Social Media: the Problem of Access', comprises an excerpt of that longer study, allowing the authors a space to explore in closer detail the issue of making social media archives accessible to researchers and students now and in the future. To do this, the paper addresses use cases that demonstrate the potential value of social media to academic social science. Furthermore, it examines how researchers and collecting institutions acquire and preserve social media data within a context of curatorial and legislative restrictions that may prove an even greater obstacle to access than any technical restrictions. Based on analysis of these obstacles, it will examine existing methods of curating and preserving social media archives, and second, make some recommendations for how collecting institutions might approach the long-term preservation of social media in a way that protects the individuals represented in the data and complies with the conditions of third party platforms. With the understanding that web-based communication technologies will continue to evolve, this paper will focus on the overarching properties of social media, analysing and comparing current methods of curation and preservation that provide sustainable solutions.

Keywords

digital preservation, social media preservation, access restrictions, data-driven research, web archives

1. INTRODUCTION

Social media platforms have become widely used spaces for communities to interact and share information across the globe. As these communities grow and the number of users increases, they leave behind a valuable cultural and historical record of life in the 21st century. In 2010 when Twitter donated their archive and on-going stream of tweets to the Library of Congress, head Librarian James H. Billington emphasised 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'.[1] As Billington indicates, data generated through social media have a number of unique attributes that make it a powerful resource for social science research. Unlike traditional data resources, like representative surveys, social media data are generated organically through popular web-based networking applications. Furthermore, the data is generated in a machine-readable format, easily converted into JSON or XML, thus facilitating complex data analytics that could lead to new discoveries about human behaviour and social interaction. The UK Data Forum strategy for 2013-2018 stresses: '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'.[2] [10] While current research still explores the full potential of this data, its value and importance has already been widely recognised. Social media data analytics have already yielded illuminating results, from detecting hate speech and its spread across digital communities to tracing political relationships between social media users and traditional news outlets.[3] As techniques for processing and querying social media data improve, the potential uses will increase exponentially.

However, in order to exploit this rich form of data, researchers must be able to access it and, furthermore, they must have the

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

[2] UK Data Forum (2013), p. 13.

[3] A number of research centres funded by the ESRC have published studies using social media data, especially Twitter. Among them include the Collaborative Online Social Media Observatory or COSMOS (http://www.cs.cf.ac.uk/cosmos),

ability to access historical data well into the future. Unfortunately, a number of obstacles inhibit access to social media data throughout its lifecycle. Access to acquire or capture, access to process and use, and access to share are all impeded by technological, curatorial, and legal and ethical roadblocks. The traditional relationship between researchers and collecting institutions, either university repositories or larger data centres like the UK Data Archive, does not accommodate the conditions of data such as that generated by social media. For example, the terms and conditions of most social media platforms are not commiserate with the demands of academic social science nor with the requirements for long-term preservation in archival systems. While social science researchers have an obligation to make their data available to other researchers in order to validate findings, Twitter and Facebook and other commercial platforms make a profit by selling access to their data, mainly to the commercial sector for consumer information. While archives serve user groups by providing access to curated data, social media platforms restrict storage options and forbid open sharing of data.

To further complicate the problem of access, social media data introduces a threat to the privacy of individuals due to its size and intrinsically linked structure. The OECD released a report on the use of new forms of data in the social sciences in which they warn policy-makers and researchers that 'the integration of multiple sources of data may increase the potential to identify individuals either directly or through deduction, based on combinations of characteristics.'[4] [7] Social media data, especially in conjunction with associated geo-spatial data, could put private individuals in a vulnerable position. Both the restrictions imposed by social media platforms and the potential for accidental disclosure of private individuals complicate the process of capturing and analysing social media data that further impact how long-term access to that data can be assured. Collecting organisations tasked with the responsibility of archiving this data – or who perceive a future need – will have to plan around these problems of access.

This paper responds to a growing need to collect and preserve culturally and historically important content generated through social media. It presents the work of two research centres using social media data who demonstrate the unique value social media data can have for social science research. The paper articulates the challenges to maintaining long-term access to this valuable data in order to define potential strategies to circumvent or mitigate those challenges. By examining cases of successful progress in preserving social media, this paper argues that preservation planning that accommodates current restrictions can make positive advances towards future access to today's social media and the rich treasury of information it holds about contemporary human society.

2. VALUE OF SOCIAL MEDIA IN DATA-DRIVEN RESEARCH

Researchers from across many disciplines have begun to make use of the rich data generated by social media, not least of which, data-driven researchers in the social sciences. New projects and initiatives have tackled resistance to using this new, unorthodox source of data to explore the potential insight and knowledge to be gained through social media analytics. The following use cases – the Collaborative Online Social Media Observatory and the Urban Big Data Centre – show a few examples of social media data in academic social science. The analysis and object lessons achieved through these use cases demonstrate the increasingly validated value of social media as digital heritage that requires immediate action to capture and preserve in order to ensure access in the future.

2.1 Collaborative Online Social Media Observatory (COSMOS)

Among the increasing number of research initiatives that have emerged around new forms of web data, the COSMOS initiative represents 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. COSMOS research projects cover a diverse range of research questions but broadly address issues of criminality, tension, and bias in digital communities. As part of the project 'Social Media and Prediction: Crime Sensing, Data Integration & Statistical Modelling', researchers at COSMOS have used Twitter data surrounding the 2011 riots in England alongside official statistics to analyse how members of the public used Twitter to organise and exchange information, a phenomenon which eluded the police's ability to gather useful intelligence.[5] [11] Another study, '"Hate" Speech and Social Media: Understanding Users, Networks and Information Flows', examines factors relating to the rise of socially disruptive content in order to forecast the spread of bias and possible violence.[6] These studies, alongside increased experience working within the restrictions of the Twitter Terms and Conditions, have enabled researchers at COSMOS to construct and test new methods and procedures for performing academic social science on social media data.

Through their published work, COSMOS researchers have presented a new model of social science research based on computational methodologies, methodologies tested in their ongoing studies. [2][5] Based on this new model, COSMOS have developed open source software for non-commercial use that will help facilitate large-scale social media data analytics. [2] 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 data such as that generated by social media.

2.2 Urban Big Data Centre (UBDC)

The Urban Big Data Centre in Glasgow is dedicated to developing new methods to analyse complex urban data in order to innovate solutions for the problems facing modern urban environments.[7] Though their work draws from a number of sources, social media data plays a role in enhancing and supplementing more traditional forms of analysis. In one on-going study, 'Integrated Multimedia City Data' (iMCD), the researchers at UBDC are collecting data from multiple sources to supplement a representative household survey in order for social scientists to gain a better '[understanding of] the complexity of decision-making in the areas covered by the survey and the possible influences of contextual factors'.[8] One

the Urban Big Data Centre (http://ubdc.ac.uk), and the Consumer Data Research Centre (http://cdrc.ac.uk).

[4] OECD (February 2013), p. 23

[5] Williams, Matthew L., et. al. (2013), p.461-481

[6] Economic and Social Research Council, Research Catalogue. http://www.esrc.ac.uk/myesrc/grants/ES.K008013.1/read

[7] Urban Big Data Centre, 'Our Vision'. http://ubdc.ac.uk/about/overview/vision-and-objectives

[8] Urban Big Data Centre (UBDC), Blog, http://ubdc.ac.uk/blog/2014/september/urban-life-captured-through-survey-sensors-and-multimedia

strand of data will include crawling Twitter and Facebook feeds for textual data as well as gathering visual data from social media and other web outlets. In another study carried out in partnership with Policy Scotland, social media data comprised the central resource and object of analysis. During the Scottish Independence Referendum in 2014, UBDC researchers collected tweets to track the flow of information and connections between Twitter users through the hashtag #indyref.[9] Analysis of the #indyref Twitter data revealed trends among users supporting the different campaigns and their relationship with more traditional news sources, such as BBC and ITV. The iMCD and #indyref projects at UBDC have seized on the opportunity to exploit the data generated by social media, data that provides insight into social interactions and human behaviours not possible through more traditional forms of data, like a representative household survey.

As demonstrated by these use cases, social media provides a valuable source of data for academic social science. The range of studies presented only in this small selection of use cases have yielded important insights into contemporary society, possibly offering solutions for better governance and policy-making. Academic research dedicated to exploring the potential of social media have begun to shape methodologies that effectively, legally, and ethically exploit the abundant information generated on a daily basis through online communities. [1] This progress, however, has encountered obstacles from both the commercial platforms who own the data to the technology required to process and index large quantities of data, which impede growth of current initiatives and the potential for longitudinal studies. To a great extent, these restrictions will continue to influence the ways social media will be captured, used, and preserved unless new relationships are established between commercial platforms and non-commercial organisations.

3. Problem of Access

As defined in the introduction, the problem of access to social media data has a number of aspects. For the vast majority of organisations, capturing social media data constitutes the primary problem. For the purposes of data analytics, researchers need access to representative samples of data with certain properties or within certain parameters, for instance, pertaining to a particular topic or created during or after a major event. The most effective and economic means of accessing this data is through an API, short of negotiating directly with a platform for greater access.[10] This means data scientists are already working with the reduced pool of social media platforms who provide a public API.[11] Though analytics can be performed on crawled social media websites, particularly for textual analysis, current solutions for web archiving and preserving web archives can be found in the study and practice of web archiving.[12] [8] The distinct problem presented by social media derives from the raw data generated by the platforms themselves – comprised of a network of databases – that can either be pulled from an API or acquired (or purchased) directly from a platform or from a third party data service. Once the researcher or institution has obtained the relatively small sample of social media data permitted through an API or through purchase, strict terms and conditions forbid many actions critical to robust digital preservation. Furthermore, the nature of large aggregates of data, such as social media data, increases the risk of accidental disclosure of individuals' identities. Lastly, and perhaps the most insurmountable problem for collecting institutions, is the problem of sharing data. Many social media platforms, like Twitter, forbid the sharing of any licensed data. This section outlines the overarching problems of access and their broad implications for researchers and collecting institutions: acquiring and capturing, processing and using, and sharing. Due to the limitations imposed by these problems, preservation planning must accommodate new circumstances while also allowing for inevitable changes in technology and platform terms and conditions.

3.1 Acquiring and Capturing

While there are a number of strategies for capturing social media – crawling a social media page with web harvester, collecting directly from an API, or possibly obtaining data directly from the platform – for the purposes of data-driven social science, the only real solutions are an API or direct acquisition. [4] While using a web crawler presents problems of its own – namely limiting capture to single pages excluding vital connections to other users and external links – it can provide a useful solution for collecting institutions, especially memory institutions, to capture the overall look and feel of a particular social media page for the purpose of preserving an important cultural and historical moment. Data-driven research however, requires large aggregates of data in order to apply research questions that seek to identify larger social and cultural patterns. To obtain data at this scale, they need access to the raw content, in formats like JSON or XML, to allow rapid computational processing. Getting this data faces a number of obstacles, not small among them are rate limits and technological capacity.

In order to derive significant patterns from social media data, researchers need substantial samples of data, varying in size depending on the parameters of a given study. However, most social media platforms that provide access to their API also restrict the amount of data that can be requested and how often through rate limits.[13] Furthermore, platforms provide only

[9] For more information about the #indyref project by UBDC and Policy Scotland, see the Policy Scotland blog. The first post on the visualization project can be found here http://policyscotland.gla.ac.uk/twitter-analysis

[10] Though some organisations have been successful in negotiating for greater access to data from social media, there are not many and this method should not form a core strategy unless an existing relationship with a social media platform exists.

[11] While many platforms provide an API for developers, including Twitter, Flickr, foursquare, Google platforms such as Google+ and Instagram, other platforms do not. Facebook, for instance, has recently closed access to its public API, restricting access to a select number of commercial organisations, as indicated by a notice published earlier this year: https://developers.facebook.com/docs/public_feed

[12] For more information about projects developing methods for using web archives for data analytics, see Ian Mulligan's 'An Infinite Archive? Developing HistoryCrawler to Explore the Internet Archive as a Historical Resource' at http://ianmilligan.ca/the-next-project and Peter Webster's 'Web archives: a new class of primary source for historians?' at http://peterwebster.me/2013/07/18/web-archives-a-new-class-of-primary-source-for-historians

[13] For more information about the rate limits of different social media platforms, see their API policies. Twitter has multiple API rate limits, but the chart in the following link shows limits for different types of requests:

limited amounts of data free of charge. Social media platforms like Twitter track the requests made through their APIs and repercussions for over-requesting could entail having privileges completely revoked.[14] These measures are understandable for commercial social media platforms – profit-driven companies – who run expensive platform infrastructures in order to sustain rapidly growing user communities. Social media platforms make money by selling valuable consumer data to commercial companies, so to ensure this market remains viable, social media platforms must restrict access to their data. These conditions, however, also restrict non-commercial use of the data, often preventing research initiatives with a limited budget from obtaining adequate test samples.

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. Similarly, third party social data providers, like Gnip or Datasift, also do not disclose their algorithm for selecting data. Though keeping methods and processes for selecting data ensures a competitive advantage for commercial entities, without the means of verifying the selection method for a data sample, academic researchers cannot fulfil the same standards applied to more traditional sources of data, like representative surveys.

One possible solution to circumventing the problem of API rate limits and undisclosed algorithms for generating data samples is to receive data directly from a social media platform. On a large scale, the only instance of this type of agreement that exists is between Twitter and the Library of Congress in the US. As mentioned in the introduction, Twitter donated its entire archive of Tweets from 2006 to 2010 and an on-going transfer of all streaming tweets to the Library of Congress in 2010. [6] This gift from Twitter will provide the Library of Congress with an important source of data for future researchers, however, ingesting the gift of tweets faced a number of problems. At the time of the donation, the Library did not have the capacity to actually transfer the data from Twitter – an expensive undertaking. To solve the problem they used a third party service to transfer the archive to the Library, awarding the contract to Gnip (now owned by Twitter).[15] [6] The Library continues to pursue the curation of this archive to make it available to researchers under particular terms but face significant technological and curatorial challenges, not the least of which involves complex IPR and data protection implications. [6] Though a direct transfer of data from a social media platform may appear to be an attractive solution, in reality, the required technological capacity and curatorial effort undermine the effectiveness and usefulness of data for researchers and collecting institutions.

3.2 Processing and Using

Besides acquiring data, the problem of access and social media also involves the use of that data once a non-commercial user has obtained it. Among the obstacles facing the use of social media data, technological capacity could present a primary issue for a large number of institutions. Depending on the volume of data required, the technological demands might be quite high, often requiring a machine more powerful than an average desktop computer. The terms and conditions issued by social media platforms also restrict the ways data can be used – including re-display requirements, storage regulations, and constraints on the use of geographical attributes (or 'geotagging') – thereby further limiting how researchers can access the data they hold. Researchers and collecting institutions also hold an ethical responsibility to protect the individuals represented in the data, protection that even anonymisation might not offer. [3] The degree to which these restrictions will affect different researchers or collecting institutions will depend on the volume of data they hold and for how long, any institution looking to support large scale – or growing small scale – data analytics will face an increasing demand for technological capacity and processing power.

Like other forms of big data, such as transactional data, social media data poses a significant challenge for storage and processing, including indexing the data for implementing search functionality. The research team at COSMOS has developed a solution to processing social media data based on the needs of their current studies. COSMOS ingests 1% of the Twitter Streaming API daily into a local NoSQL database where they store the raw JSON data. After three years of collecting at this rate, the COMSOS archive currently holds somewhere in the vicinity of 2 billion tweets. After collection, the team then add an autonomous layer of added attributes – such as location, gender, age, occupation – derived from the JSON data. With this infrastructure, the researchers at COSMOS can more easily perform analytics for a variety of research questions. As their archive of social media data increases, however, they will have to continue to scale up to new technological solutions. For most institutions interested in collecting a large archive of data for on-going studies using social media data, this structure of databases may provide a helpful solution, however, further external restrictions will continue to limit their use of the data.

Once a researcher or collecting institution has acquired data from a social media platform, they must operate within the contractual confines of the social media terms and conditions. The Twitter Developer Agreement and Policy, for instance, forbid users 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'.[16] 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 cloud storage provider. Twitter is not unique in licensing their data under a non-transferrable agreement. Cloud storage could have been an effective solution for many organisations facing long-term storage needs for their social media archives who do not have the budget to build a local storage facility.

Use of archived social media data also poses potential threats to privacy and data protection, in ways that use of traditional source do not. As referenced in the introduction, the intrinsically linked nature of big digital data makes it easier to accidentally disclose the identities of private individuals. When multiple sets of data, including administrative data and transactional data, are

https://dev.twitter.com/rest/public/rate-limits. For more general comparisons, the public website API 4 DEV offers a useful comparison of different free API rate limits: http://www.api4dev.com.

[14] Twitter Developer Agreement and Policy (2015). https://dev.twitter.com/overview/terms/agreement-and-policy.

[15] Twitter blog post by Jana Messerschmidt, 'Twitter Welcomes Gnip to the Flock' (April 2014). https://blog.twitter.com/2014/twitter-welcomes-gnip-to-the-flock.

[16] Ibid. Twitter, Developer Agreement (2015).

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 anonymisation might not fully prevent such accidental disclosure. In a report to the White House last year, 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'.[17] [3] These conditions make it imperative for researchers to adhere to new ethical standards and precautions when using social media data.[18]

3.3 Sharing

The short answer when considering the potential for sharing social media archives between institutions is 'no'. Ultimately, social media platforms forbid sharing access to licensed data. A few exceptions may exist to this condition, but only in extraordinary circumstances, such as the donation of the Twitter archive to the Library of Congress, who will be subject to a range of other contractual restrictions once they open access to their archive. [6] Furthermore, sharing social media archives, even for non-commercial purposes, poses other legal risks, such as infringing intellectual property rights, violating data protection laws, and possibly even violating an individual's right to be forgotten.[19] As previously discussed in the problem of using data, sharing social media archives could also increase the chances of accidentally disclosing the identity of individuals. The problem of sharing data – making it accessible to people outside the named individuals on a license agreement – may pose the greatest problem of access for researchers and collecting institutions, whose objective is to publish or circulate valuable content for non-commercial use – for validating scientific conclusions, for academic research, for education, and for cultural and historical memory.

For data-driven researchers like those at COSMOS and UBDC, 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 organisation who 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 their data.[20] 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 and 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'.[21] 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 social scientists from sharing their data in a meaningful way without violating social media platform terms and conditions.

While rights issues stemming from platform terms and conditions are the principal rights restrictions facing researchers and their partner collecting institutions, other potential legal and ethical issues also impede the sharing of social media data. The intellectual property rights embedded within data such as social media, from text to image to sound to moving image, introduce a profusion of restrictions to how social media data archives could be accessed or re-displayed. As mentioned earlier, the scale of data introduced with sources like social media platforms increases the risk of accidentally disclosing the identity of private individuals. This potential risk might require measures beyond those already used by institutions who hold digital content containing personal information. Furthermore, the technology required to process large aggregates of data by user request also limits organisations to the amount of data they can reasonably maintain. The Library of Congress has been coping with this reality since 2010 when they received the Twitter donation. [6] The Library has yet to develop or acquire the technological capacity to make their Twitter data accessible to researchers five years on.[22]

Though difficult, the problems that face the capture, use, and sharing of social media data are not insurmountable. As platforms like Twitter take the first steps in negotiating on broad terms with non-commercial organisations, social media companies may discover value in improving their public image through providing data to non-commercial institutions. The potential value of social media data, as presented in its early stages at COSMOS and UBDC, should not be underestimated – it warrants the compromises and continued effort to overcome access restrictions in order to one day open this rich data source to researchers and future generations.

4. SOLUTIONS FOR LONG-TERM ACCESS TO SOCIAL MEDIA

The capture, use, and long-term preservation of social media, as demonstrated, poses difficult challenges for data-driven researchers and for collecting institutions. However, a number

[17] Executive Office of the President, p. xi

[18] Housley, p. 175.

[19] EU rulings on the right for individuals to have their personal information removed from internet search engines in certain circumstances has a significant impact on the practices of organisations working with digital content sourced from the web. For an overview of these implications for big data, see 'Forgetting Footprints, Shunning Shadows. A Critical Analysis of the "Right To Be Forgotten" In Big Data Practice' (2011) by Bert-Jaap Koops at http://script-ed.org/wp-content/uploads/2011/12/koops.pdf or more general information about the Right to Be Forgotten, see the European Commission Factsheet at http://ec.europa.eu/justice/data-protection/files/factsheets/factsheet_data_protection_en.pdf.

[20] For a list of data funder policies, DCC has a compiled list on their website: http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies.

[21] Ibid. Twitter, Developer Agreement (2015). Guiding Principles Section 6(b).

[22] The Library has not yet opened their Twitter archive to the public but they have released a limited amount of data to a select number of researchers through their data grants. More information about Twitter data grants can be found on their blog post 'Introducing Twitter Data Grants' from 5 February 2014 at https://blog.twitter.com/2014/introducing-twitter-data-grants.

institutions have managed to successfully archive social media and have taken steps to ensure its long-term accessibility. As described above, the team of researchers at COSMOS have developed a system for ingesting and processing Twitter data in a way that will support long-term preservation. In the sphere of collecting institutions, while national memory institutions such as the British Library and the National Library of Scotland archive social media as harvested websites, the UK Government Web Archive has pursued another strategy by capturing the Twitter data for the public tweets issued by official central government Twitter accounts.[2324] At the publication level, a number of smaller initiatives have made progress in the use and re-use of social media data as well through promoting good practice in the individual use cases of social media data analytics. These solutions may not solve all the challenges posed by social media, but they provide a framework for moving forward.

Both researchers and collecting institutions have begun to make progress in the curation and storage of social media. In terms of sharing, the UK Government Web Archive at The National Archives have captured tweets from targeted accounts – official, public central government Twitter accounts – as part of the pilot stages of a two-year project.[25] The TNA Twitter archive excludes re-tweets and other content not directly published by central government and they clearly state limitations on re-use on their website.[26] However, the raw JSON and XML data available on their website make the Twitter data accessible for computational analytics. Though both the researchers at COSMOS and TNA have achieved a level of success in archiving social media, both solutions also have limitations. COSMOS is not, for instance, able to share their archive of tweets with other researchers. At TNA, they capture only central government tweets, thereby limiting the archive's broader use for data analytics.

As the larger research and collecting institutions make progress in the preservation of social media data, researchers and archivists can proceed with other strategies for facilitating the on-going use of social media. Some data-driven researchers, for instance, have started publishing analysis with a statement about the availability of the data used during the study. In a recent article by COSMOS, 'Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class form Twitter User Meta-Data', the authors have issued a 'Data Availability Statement' on the title page to explain their process of acquiring the data and processing it for the required results and the level to which they can share the data. [9] This type of statement is one way to acknowledge that social media data cannot be used and shared the way traditional data sources might be, thus, over time, facilitating collaboration across institutions to agree on methods for publishing and curating social media data. This cooperation could help meet one of the major challenges defined in the OECD report on 'New Data for Understanding the Human Condition'.[27] [7] In this report, the OECD define the lack of access to data sources as an obstacle to data sharing, and thus an obstacle to effective social science research. As a solution, they encourage cooperation across institutions and governments internationally. Identifying the availability of data used in academic publications as well as within archived collections facilitates cooperative tactics and also makes a demonstration of compliance with social media terms and conditions.

The current relationship between social media platforms and research and collecting institutions depends on trust and mutual agreements. In order to facilitate further negotiations directly with platforms, it is important for non-commercial institutions to assert their trustworthiness and their dedication to compliance with the terms and conditions established by the owners of the data. Trustworthiness and reliability, as well as relationship-building with depositors, is not a new activity for collecting institutions; this new phase in the creation of digital information brings a new generation of producers to assure of the integrity and security of digital repositories.

5. CONCLUSION

Social media platforms, like any web 2.0 application, behave differently from the information web and pose different challenges to long-term preservation. The data generated by social media lends itself to research questions asked by social scientists because it moves quickly, providing real-time reactions to major events in both the terrestrial and digital world. In order to create strategies for archiving social media, collecting institutions will have to build close relationships with research institutions using social media data to understand the likely needs of those who will use the data in the future. With the increase in born digital content, collecting institutions have already successfully adapted their practices and are equipped to evolve to meet the demands of new forms of digital information like social media. As the size of digital content grows, an even greater need arises for strict retention planning and careful consideration of selection policies for collecting institutions. Collaboration with other institutions can help individual organisations make informed decisions about what to capture and how to curate it. Collaborating across institutions and national governments can also help overcome the challenges to capturing, using, and sharing social media data through providing coordinated negotiation with platforms. Through collaborative progress, research and collecting institutions can build a community where researchers and user communities can share tools, methodologies, and resources.

6. REFERENCES

[1] Burnap, Peter, Rana, Omer, and Avis, Nick, 'Making Sense of Self Reported Socially Significant Data Using Computational Methods', International Journal of Social Research Methodology, Computational Social Science: Research Strategies, Design and Methods. Volume 16, 3 (2013).

[2] Burnap, Peter, Rana, Omer, Williams, Matthew, Housley, William, et. al., 'COSMOS: Towards an Integrated and

[23] The Legal Deposit Libraries (LDLs) in the UK archive the web through non-print legal deposit and through a permissions-based collecting policy. The non-print legal deposit web archive can only be viewed on the premises of one of the LDLs. The permissions-based archive, the UK Web Archive, can be accessed online at http://www.webarchive.org.uk/ukwa.

[24] The National Archives, Records, 'UK Government Web Archive: Twitter'. http://www.nationalarchives.gov.uk/webarchive/twitter.htm#govOrg.

[25] The National Archives, Information on web archiving, 'Using the social media archive'. http://www.nationalarchives.gov.uk/webarchive/information.htm#using-the-social-media-archive.

[26] The National Archives, Information on web archiving, 'Re-use of content accessible through the UK Government Web Archive'. http://www.nationalarchives.gov.uk/webarchive/information.htm#using-the-social-media-archive.

[27] OECD, p. 25-6

Scalable Service for Analyzing Social Media on Demand' (2014). DOI= http://dx.doi.org/10.1080/17445760.2014.902057.

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

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

[5] Housley, William, Williams, Matthew L., et. al. (Eds.) 'Computational Social Science: Research Strategies, Design and Methods', International Journal of Social Research Methodology Special Issue, 16, 2 (2013).

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

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

[8] Pennock, Maureen. 'Web-archiving'. DPC Technology Watch Report (March 2013). DOI= http://dx.doi.org/10.7207/twr13-01.

[9] Sloan, Luke, Morgan, Jeffrey, Burnap, Peter, Williams, Matthew, 'Who Tweets? Deriving the Demographic Characteristics of Age, Occupation and Social Class from Twitter User Meta-Data' (2015) PLoS ONE 10(3): e0115545. DOI= http://dx.doi.org/10.1371/journal.pone.0115545.

[10] UK Data Forum, 'UK Strategy for Data Resources for Social and Economic Research' (2013), p. 13. http://www.esrc.ac.uk/_images/UKDF-strategy-data-resources_tcm8-26806.pdf.

[11] Williams, Matthew L., Edwards, Adam, Housley, William, et. al. (2013) 'Policing cyber-neighbourhoods: tension monitoring and social media networks', Policing and Society: An International Journal of Research and Policy, 23:4, 461-481. DOI= http://dx.doi.org/10.1080/10439463.2013.780225.