Ashlyn Velte. Activist Social Media Archiving: Practices, Challenges, and Opportunities. A Master's Paper for the M.S. in L.S. degree. April, 2016. 84 pages. Advisor: Stewart Varner

Social media has played a significant role in recent activist movements. It empowers activists to organize and communicate their experiences. Archival efforts to document narratives that are historically silenced makes activist material an attractive collecting area. However, archives trying to preserve digital ephemera like social media from activist movements face digital preservation challenges as well as ethical considerations. By conducting online surveys and semi-structured interviews with archivists working on projects collecting activist social media this study found that activist social collecting projects: 1) face ethical and collection development challenges, 2) usually follow traditional models for acquisition, description, and access, and 3) increase donor and user engagement with collections. This suggests that in the future the profession would benefit from the development of best practices surrounding ethical collection development and use of activist social media.

Headings:

Web Archives

Appraisal of archival material

Social Media

Digital Preservation

Metadata

Archives—Access Control

ACTIVIST SOCIAL MEDIA ARCHIVING: PRACTICES, CHALLENGES, AND OPPORTUNITIES

by
Ashlyn Velte

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Library Science.

Chapel Hill, North Carolina

April 2016

Approved by

_____

Stewart Varner

Table of Contents

Introduction

The last several years have seen an increase in activist activity within the United States. For activist groups online digital tools, like social media, play an important role in proliferating and spreading their organizing efforts. Without social media it would have been much more difficult for recent police violence protests to garner recognition by the main stream. Kimberly Springer (2015) says that it is "indisputable" that social media played a significant role in bringing police violence against African Americans into recent public consciousness. Indeed, Bergis Jules (2014) confirms that leaders of the Black Lives Matter movement against police violence value social media platforms in "organizing and curating the Ferguson story." Some academic research has started exploring links between social media and social movements, or "protest networks" as W. Lance Bennett and Alexandra Segerberg (2011) call them.

Jules (2015) writes about how the storm of social media posts during the Ferguson protests allows researchers and archivists access to current public sentiment on a scale that was not possible before. He says that "the ephemeral nature of this new type of record challenges us to develop tools and strategies to capture and preserve that digital content for future use." Importantly, he explains how the narrative of Ferguson would differ without the social media record; social media allow events to be seen from the perspective of the protesters themselves, instead of relying solely on images and video captured by mainstream media. Since social media plays an important role in today's

social movements, then it naturally falls under the collecting scopes of many cultural heritage institutions, such as libraries, archives, and museums. Particularly since social media usually reflects public sentiments on significant events, then it is important for these collecting institutions to preserve social media for researchers interested in today's events. However, the relative newness of the format makes it daunting work for some institutions since there is a lack of consensus on best practices for activist social media.

Despite its apparent newness, there are several widely discussed projects that test the waters for collections of social media from activist events and organizations. For instance, there are at least three different digital projects collecting digital ephemera from the Occupy Wall Street Project. Emory Libraries' Archive of Occupy Wall Street Tweets contains over 10 million tweets documenting the movement against economic inequality in New York City. Its aim was to provide data visualization to help researchers understand the role social media plays in social movements (King, 2012). The Occupy movement also has its own digital archive hosted by volunteers at George Mason University and the Roy Rosenzweig Center for History and New Media (Erde, 2014). Though not specifically focused on social media, their archive maintains the "digital evidence and stories from the Occupy project" (Roy Rosenzweig Center…, 2011).

More recently, similar digital archives have results from the social movement surrounding the Ferguson protests after the death of Michael Brown in August 2014, who was shot by a white police officer in Ferguson, Missouri (Buck et al., 2014). Like the Occupy movement, several digital archives document this movement that exists primarily online. However, new to the Ferguson protests was the notion that civilian journalists via social media could publicize police violence not covered by mass media outlets. The

Documenting Ferguson project at Washington University in St. Louis provides an online access and donation portal for digital content related to the Ferguson protests. This digital archive partnered with other institutions in the St. Louis area to be sure it was adequately representing the need of the community (Buck et al., 2014). Other archives focused specifically on the social media content generated by people involved in the protests. Ed Summers at the Maryland Institute of Technology in the Humanities has harvested a data set of Ferguson related hashtags from Twitter (Jules, 2015).

After the death of Michael Brown a series of protests erupted nationwide against acts of police violence. One of the more heavily covered protests occurred in Baltimore after the death of Freddie Gray on April 12, 2015. During the protests, the city of Baltimore called in support from the National Guard. The Maryland Historical Society, the University of Baltimore, and the University of Maryland Baltimore County created an Omeka site (baltimoreuprising2015.org) where users could directly contribute their social media content to help tell the story. The Maryland Historical Society says that they "believe that it is important that the materials are preserved in a historical repository so they are not lost" (Maryland Historical Society, 2015). They already have over forty individuals who have contributed their collections to the site.

In addition to these national protests, the last few years has seen an increase in activism on college campuses. During the fall of 2015, several campuses made national and international news when students held events demonstrating against the lack of diversity in college environments. Protests at Yale University, the University of Missouri, Ithaca College, and Claremont McKenna College resulted in real recognition and action taken by the administration. Ithaca College appointed a diversity officer, and the

president agreed to step down in the summer of 2017 (Maycon, 2016), and at Claremont McKenna College the dean of students stepped down. ("Student Activists Nationwide…", 2015). This increase in activism around campus diversity has generated conversations about the state of racial equality in U.S. Several large institutions with well-established archives started collecting initiatives around student activism (e.g. Drake, 2015; Farrell, 2016; Kolachina, 2015). Developing these collecting areas has been at least partially inspired by the increase in the activity seen on campuses.

These examples of activist organizing from the last 10 years coincide with a significant social media presence behind their causes. Additionally, their activities have significant impact on the conversations happening in the media and the government. If significant evidence of these events occurs online then libraries, museums, and archives have a role in preserving it for future generations. Without archives of the Occupy movement, Ferguson protests, and the Baltimore Uprising, the digital ephemera related to the events would have disappeared. Their materials have pushed the bounds of what archives traditionally a record; digital content, especially social media content, is not traditionally collected by repositories, and these institutions have not established best practices for this new form of digital media. Significant risk for loss of this materials has entered into the professional awareness.

In recognizing the risk for loss Jeffrey (2012) identifies the potential for a second 'digital dark age' in regard to social media. The first Digital Dark Age, he claims, occurred with the widespread adoption of computing systems without the proper procedures for preserving the records they created. He claims that this material was lost due to the obsolescence of the software and media, data corruption as media degrades,

and inadequate metadata. Though standards and systems addressing these problems have been identified, he claims that similar problems currently face newer online technologies like social media. The risk of loss for online materials has entered mainstream consciousness as both The Gaurdian and NPR published stories about the risk of data loss from new technologies (Sample, 2015; Morning Edition, 2016).

Loss of online records is a great risk for activist groups who typically have limited power within the social structure. Kimberly Springer (2015) notes the importance that ephemeral records play in documenting cycles of contention (where people realize that the problems they face are systematic rather than isolated incidents.) For activists and organizing groups archives serve to explain their work without relying on the mainstream media or politics which sometimes "criminalize and capitalize on dissent" (Springer, 2015) She cites work by Howard Besser (and his activist-archivists.org) that explains to activists that archives provide accountability for those in power, accessibility to their records, self-determination in defining their social movements, education materials for tomorrows classes, and continuity through lessons for future activism.

If the risk of loss has higher stakes for activist organizations, archivists must understand what is currently being done to preserve online activism so that we know what works and what can be improved. This research study seeks to examine the current practices of some of the institutions collecting activist social media in order to identify common practices, challenges, concerns, and outcomes. In order to suggest possible areas of future growth for activist social media archives, current practices must first be identified.

Literature Review

Understanding where activist social media archives are situated in the professional literature involves reviewing a complicated interplay of archival theories and practices. It first means defining archives' role in building historical narratives, by explaining defining archival silences and collection strategies to overcome them. Primarily, archival collections turn to community archives to fill in gaps in the historical narrative. Activist collections, as another type of community archive, are now also being used to address missing histories. The examples in the introduction demonstrate that an increasing number of activist records are being created online. In order to understand how cultural memory institutions might best engage with born-digital content from activist groups, we must also review digital preservation and web archiving strategies. Lessons from community archiving and digital preservation impact the way that archives approach and collect both analog and digital materials from activists.

1.1 Archival Silences

Activist collections share similarities to community archives, particularly in that they are intended to address archive silences. First though it is important to understand the way archives have perceived collection development over time. Archives provide evidence of history. This 'evidence' might refer to legal evidence, but it more commonly refers to documenting collective memories and communicating information (Millar,

2010). Newcomers to archives often arrive with the assumption that the preserved collections accurately reflect historical reality. Discussions of archives and archivists' role over the past fifty years indicate that the archivists' traditional viewpoint was that they were the neutral documenters of the past. Cook (2013) names four archival paradigms that describe the dominant thought processes of the profession throughout the history. He defines his paradigms as professional *"frameworks* for thinking about archives, or archival mindsets." (p. 97). The earliest paradigm he identifies, "evidence" refers to a period in the late 19th and early 20th century, where archives were perceived as the primary way to guard "truth." Despite archives having progressed through three other of Cook's (2013) paradigms, this traditional view of archival records persists within areas archival thinking. Other authors have remarked that many consider archives to be a place for objective recordkeeping. Indeed, Verne Harris (2002) says that as "many archivists are wont to argue, the repositories of archives are the world's central memory institutions… the idea is that archives reflect reality" (p. 65).

However in the 1970s archivists began discussing the possibility that archives are not the bastions of objective truth that they were assumed. Archivists began to understand the role they play in shaping historical narratives when Howard Zinn (1970) asked archivists to step outside of the traditional collecting scope. Controversially for the time, he recognized that, despite archivists believing they exist behind a wall of neutrality, collecting decisions often support dominant power structures. Specifically, archives collect from rich and powerful groups, such as military, political, and business leaders leaving ordinary people out of the archival records. To offset the historical bias of existing collections he suggested that archives strive to document common experience.

Five years later, Ham (1975) gave the presidential address to the SAA. Following Zinn's (1970) shattering claims on archival neutrality, Ham (1975) made a call for the profession to live on the "archival edge," outside of mainstream culture to better document the breadth of human experience rather than limiting the documented past to mainstream culture. This means that traditional passive collecting with a limited view on the definition of what constitutes a record will not accomplish what archives have always tried to do: "hold up a mirror to mankind" (p. 13). He says that if archivists do not reflect reality human kind they cannot learn from the past.

Verne Harris (2002) provides a powerful (yet extreme) example demonstrating how archives fail to accurately reflect reality. He describes South Africa's apartheid record keeping as anything but "objective" and how the State Archive Service (SAS) increased the power of the apartheid system. The government itself was reliant upon "control over social memory, a control which involved both remembering and forgetting" to achieve dominance over the groups it oppressed (p. 69). The apartheid government achieved social memory construction because the SAS was part of the state functions. This means the SAS did not reach out to underserved and uneducated populations, facilitated government objectives through record keeping, and also was poorly placed to resist the government. Additionally, several government departments refused to transfer records for public accessibility. Many apartheid records were destroyed between 1990 and 1994 despite SAS intervention. All of this supports the main concept Harris (2002) introduces, which is that archives can only save and show posterity a small portion of the past. The small amount of evidence that gets preserved is what he calls an "archival sliver." Further, archivists actively construct this sliver. The National Archives of South

Africa, created after the end of apartheid, recognize the active role archivists play in memory construction. Recognizing that archives are situated within existing power structures, they seek records from previously undocumented groups.

The gaps in the historical record created consciously or unconsciously by the dominant group (the group that has the most social, political, or economic power) are known as archival silences (Carter, 2006, p. 217). Carter (2006) explains that groups silenced by archives are prevented from participating in historical and societal dialogues. As he explains "archivists are constantly confronted with choices about what to include and what to exclude… Limited resources and/or a lack of understanding ensure that all records are not given equal attention" (p. 219). To Carter (2006) silence results in the loss of societal memory, failure to form a collective identity particularly for the marginalized, and produces further victimization.

The third of Cook's (2013) archival paradigms was a search for professional identity. During this search, archivists recognized the pluralistic nature of records where one historical perspective does not capture 'reality.' Though his discussion of this paradigm mostly focused on building the identity of the profession by identifying areas of expertise and common values, this paradigm also involved recognizing that archivists are influenced by dominant groups within society. Meaning that the profession considered part of its identity as being unwittingly responsible for shaping societal narratives for history and memory.

If archivists build social memory as Harris (2002) and Carter (2006) suggest, what happens if they strive for neutrality? Some have suggested that neutral libraries may in fact do more harm than good. Pagowsky and Wallace (2015) explain that neutral

libraries passively support existing oppressive power structures by failing to recognize the different experiences of their patrons. They provided support and resources on the Ferguson protests and institutional racism which positively impacted their patrons. Their "Black Lives Matter" library guide has received much campus support, and has deepened ties between faculty, students and the library on their campus (University of Arizona). They explain that libraries have an obligation to provide support to underrepresented groups rather than remain "neutral" and silent. Compellingly, they claim "Systemic racism can't be confronted or resolved unless *everyone* is involved in its interrogation" (p. 199). Their argument adds importance and immediacy for archives to document those existing on Ham's (1975) archival edge. As archivists recognize how easy it is to miss societies' pluralism, seek ways to actively collect from groups that are historically underrepresented in society.

The recent focus on community archiving is the most recent archival paradigm experienced by the profession (Cook, 2013). Archivists' working for and within community archives addresses the problems first articulated by Zinn (1970) and Ham (1975). Indeed, Flinn (2007) says that community archives provide a place for archivists to "leave behind the idea of archivist as a neutral, passive, reactive figure and instead embraces a much more active or proactive role, one which acknowledge the power and influence which the archives has over framing our archival heritage and social memory" (p. 168).

Defining community archives can be challenging because of the immense differences between each community, and because community identities sometimes change quickly (Flinn, 2007). Indeed, Bastian and Alexander (2009) explain that "there is

no one definition of community" as each community differs depending on context (p. xxii). Some communities might be based on locality, on identity, or on shared values and beliefs (Flinn, 2007). Because of these challenges inclusive and broad definition of 'community' work best. For instance a community is a "group who define *themselves* on the basis of locality, culture, faith, background or other shared identity or interest" (Flinn, 2007, p. 153). For the same reasons, defining a community archive is also difficult. To Flinn, Stevens, and Shepard (2009) community archives are "collections of material gathered primarily by members of a given community and over whose use community members exercise some level control" (p. 73). The process of gathering material is largely a grassroots effort, meaning that much of the work begins, and often remains, outside of established institutions (Flinn, 2007).

Because community groups often have a common sense of place or locality it might be initially difficult to differentiate them local historical societies. The defining difference is that among community archives' motivations arise from a need to document the marginalized; existing power imbalances inform both their identities and their decisions about what and how to preserve those identities (Caswell, 2014). Despite immense differences there are some characteristics shared between community archives. Caswell (2014) identifies several common principles of community archives from both existing literature and her experience. They provide examples of how community archiving differs from traditional models of archiving. Caswell's (2014) community archiving principles are participation, shared stewardship, reflexivity, multiplicity, and activism. Each of these principles informs this study in some way so I will spend some

time describing them. However, two of these principles most directly relate to this study, the principles of multiplicity and activism, which I will elaborate on more fully.

The first principle, participation, is an integral concept for community archives. This principal refers to the active involvement communities have in all parts of the archival process from appraisal to description to creating access restrictions (Caswell, 2014). The success of community archives largely hinges on the active involvement of the community members. This principle presents a shift from a paradigm where archivists are experts to one where archivists are collaborators (Cook, 2013). This is important to communities because active participation allows them to define their own history (Flinn et al., 2009). The second principle relates to participation in that it requires community engagement. The principle of shared stewardship diverges most clearly from the traditional archival model. According to dominant archival practices, established repositories assume ownership of materials, giving them total responsibility for their maintenance and use. However, for communities working with established repositories "the community maintains some ongoing autonomy over the records that originated within it" (Caswell, 2014, p. 312). This principle is further elaborated on by Flinn et al. (2010), who explain that partnerships between repositories and communities should anticipate long-term commitment to maintain the materials together over time. The principle of reflexivity, or self-reflection, relates to building the story told by an archive. Self-reflection by both individuals and the group at large helps community archives most effectively tell their story, and react to the often contradictory viewpoints present within the group.

Caswell's (2014) next principle, multiplicity, refers to two different aspects of community archives. First, community members rarely agree all the time. This means that there are many different and conflicting perspectives among members of a community. Thus community archives do not just capture commonalities within the group but also its differences. The acknowledgement of diversity can be liberating for communities that have been viewed by the mainstream as one clearly defined group (Caswell, 2014). Having diverse viewpoints then is a shared characteristic across many community archives. The other way in which community archives have multiplicity is in format. For instance, most community archives push the bounds of what has been traditionally defined as a record (Bastion & Alexander, 2009; Flinn, et al., 2010; Caswell, 2014). Preserving previously undocumented kinds of material means that the archive better reflects the communities they originate in. However, for mainstream repositories this multiplicity of format can introduce preservation challenges.

Caswell's (2014) principle of activism is the most relevant to the current study, and the one in which I will elaborate on most fully. This principle identifies that for communities, documenting their history is way of pushing against mainstream erasure. Caswell (2014) says "the creation of community archives can be seen as a form of political protest that it is an attempt to seize the means by which history is written and correct or amend dominant stories about the past" (p. 314). She identified this principle of community archives by recognizing how across the literature mentions of what motivated their creation refers to social change and increased mainstream visibility of a group. For instance, Flinn and Stevens (2009) argue that independent archives, or community archives that remain separate from large collecting institutions, may in fact be social

movements. They serve as a way to challenge dominant understandings of heritage and also act as a way for groups to create their own identities. Since mainstream institutions often fail to collect marginalized stories, independent archives act as a way to subvert dominant historical narratives; they seek to build a more complex social understanding of history. Similarly, Flinn, et al. (2010) discuss community archives as a way to share voice among communities that otherwise are not reflected in official records. These characteristics of community archiving should inform archival projects that document activist groups.

1.2 Digital Preservation

Based on recognition of and efforts to reduce archival silences, perhaps it is no surprise that archivists turned their attention to recent activist movements. Collecting from activist or advocacy groups overlaps with community archives as demonstrated by Caswell (2014). Yet, recent records of activism include an increasing amount of born-digital materials from the open web; activists use social media to organize and communicate their message. This means that archivists face the challenge of preserving this born-digital content.

Born-digital material, is content that is "created and managed in digital form" (Erway, 2010, p. 1), meaning that it never had an analog incarnation during its use. Social media, also called Web 2.0 or social networking, are born-digital records of large scale communication between communities and individuals on the internet. Web 2.0 applications allow users to interact with and generate online content (Zeng et al., 2010), where before users could only view static information provided on webpages. There are many different web platforms that allow users to generate and interact with each other's

content. There are platforms for video sharing (such as YouTube and Vimeo), platforms for photo sharing (Flickr and Instagram), platforms for social networking (Facebook and LinkedIn), platforms for blogging (WordPress and Tumblr), and for microblogging (Twitter) (Jeffrey, 2012).

Digital media changes faster than archivists and researchers can create tools to adequately preserve it. As mentioned earlier, Jeffrey (2012) describes that a lack of methods to preserve early digital documents led to a 'Digital Dark Age,' where the rapid change in technology led to 1) data corruption, 2) media and software obsolescence, and 3) inadequate metadata. Social media, as a form of digital material, also faces these risks. Preserving social media content involves implementing digital preservation best practices to combat these risks. Archivists have developed methodologies and practices to protect against each of the risks that Jeffrey (2012) identifies which also apply to the preservation of social media.

Firstly, digital preservationists identify ways to combat both data corruption and media obsolescence. Digital material has specific preservation concerns like bit rot, obsolescence, and authenticity (Erway, 2010). The OCLC have resources documenting steps a digital repository should implement to preserve digital content including tools and procedures used to combat these issues (Erway, 2010). For example, obsolescence can be managed by creating specific metadata regarding how a file should be read, while authenticity can be assured by installing write-blockers and creating checksums for the files (Barrera-Gomez & Erway, 2013). These practices should also apply to the management and preservation of social media data, but the unique issues associated with social media harvesting and access make it difficult to comply with these suggestions.

Social media corruption and obsolescence occurs because content changes quickly and is susceptible to loss. Social media loss could occur for two reasons. Firstly, loss might occur due to the commercial nature of the platforms on which the data lives (Jeffrey, 2012, p. 560). For example, Flickr once accidentally lost 4,000 photographs (Wauters, 2011, cited by Jeffrey, 2012). In another case the platform of GeoCities shutdown without a preservation plan, and resulted in the loss of a massive amount of information (Thomson, 2016, p. 17). Another reason loss might occur is because a user or some other third party removes or edits posts. This data might be important to preserve, but before an archive decides to maintain deleted social media content a repository may need to contend the ethical issues regarding privacy described below. For instance, Twitter has policies in place to protect the privacy of users who have deleted Tweets (p. 21).

Digital preservation strategies also identify suggestions for the creation of adequate metadata. Metadata is important for the proper preservation of digital objects. It is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (NISO, 2004, p. 1). It usually contains the aggregate knowledge about any information object, which usually consists of its content (information about what the object contains), context (related information outside of an objects apparent content), and structure (information about how an object's formal relationships to itself or to other objects; Gilliland, 2008, p. 2). The purpose of good metadata is to facilitate discovery, organization, interoperability, digital identification, and preservation of resources (NISO, 2004, p. 1-2). Different metadata standards serve

one or more of these specific purposes. Archives and libraries use several common metadata standards to describe born-digital material, such as:

- **Dublin Core.** The Dublin Core Metadata Element Set is meant to be a metadata scheme that could apply to any web based resource. It provides a simple set of elements that can be applied by both cataloguers and non-cataloguer. Its element set is simple and concise, and each element is optional or repeatable. It is used by those both in and outside of the library community (NISO, 2004, p. 3).
- **Metadata Encoding and Transmission Standard (METS).** METS is an XML schema for defining the structure and description of objects in a digital library. It includes technical metadata about the digital objects to ensure the interoperability in case of future migration. It is more technical and more extensive than Dublin Core (NISO, 2004, p. 4-5).
- **Metadata Object Description Standard (MODS).** MODS is also an XML schema based on MARC 21 records. It is hierarchical and descriptive and also more detailed than METS. For this reason it works well in concert with the structural elements of METS (NISO, 2004, p. 5-6).
- **Encoded Archival Description (EAD).** EAD is an XML structural schema for elements of archival finding aids. It allows for easy online display and searching (NISO, 2004, p. 6).
- **Describing Archives: A Content Standard (DACS).** DACS is the official content standard for describing archives in the U.S. Most commonly used in finding aids to describe the content and context of a collection (Society of American Archivists, 2013).

These practices to preserve born-digital content also apply to web archives and social media archives. However, social media introduces format specific preservation practices and challenges.

1.3 Web Archiving and Social Media

Web archiving is the process of collecting and preserving portions of the World Wide Web in an archival format for future use (International Internet Preservation Consortium, 2012). Web archives are a specific type of born-digital record, where the data generated by harvesting web content is preserved following best practices described in the previous section. This usually means using specific URLs to capture web pages and display them to users as they existed at the time that they were captured (as with the Internet Archive's Wayback Machine). Yet, as a relatively new practice within digital archiving, there remain a few determined best practices.

In an environmental scan published as a Harvard Library report, it is explained that many research libraries "recognize website archiving ("web archiving") as an essential component of their collecting practices" (Truman, 2016, p.5) By conducting interviews with 23 institutions from around the world, Truman (2016) determined that there was a wide variety of practices in place for web archiving. For instance the amount of staff dedicated to web archiving varied widely from only part of one staff member's time to up to 20 people. Yet most respondents dedicate the equivalent of around one full-time staff member to web archiving activity. Archives vary in whether they outsource the technical infrastructure to harvest and preserve websites. Some institutions use their own tools crawl and host archived websites, some outsource these function, but a majority use some combination of internal and external support for technical infrastructure. Eleven

respondents used Archive-It (the Internet Archive's contract service for web archiving) to crawl and host their web archives. Additionally, the tools used for web archiving varied greatly. Many tools have been built for granular processes of web archiving related to any stage of the information lifecycle. Truman (2016) stopped counting the variety of tools after reaching 77.

The report identified 22 ways that archivists can move web archiving forward. Yet, some of the overarching themes signify the present trends in web archiving. Specifically, that institutions find it difficult to choose what to include in their web collections because they worry about the potential for duplication of effort between collections. The report suggests better collaboration, and communication about collection development policies. In addition, there is concern about the reliance on third party vendors to host archived web content. Some of the interviewed archivists fear that reliance on vendors does not guarantee the longevity of the data. Some institutions combat this concern by downloading .warc files from a service like Archive-It and storing them on their in their local digital repositories. Truman (2016) suggests partnering with other institutions as well as these service providers for consistent access and reduce reliance on outsourcing.

In many ways collecting social media is not all that different from web archiving generally. Yet there are a few things about social media, and activist social media in particular, that introduce additional considerations, giving some archivists reason to pause when they want to begin collecting activist social media. One of the primary differences between web archiving as a whole and social media archiving specifically is the potential to harvest social media content as datasets rather than as single documents.

Social media generates large amounts of data that can be useful for researchers across a variety of disciplines. As a report on preserving social media published by the Digital Preservation coalition explains, social media data sets "will be of use on a large scale to data-driven researchers in the social sciences and other disciplines" (Thomson, 2016, p. 4). The Social Media Archiving Toolkit released by North Carolina State University Libraries identifies several research studies where large scale social media was used (2015a). For instance research using social media data has been conducted in fields as disparate as medicine (Young, Rivers, & Lewis, 2014), social science (Collins et al., 2014), linguistics (Squires, 2014) and economics (Haustein et al., 2014). Recognizing the importance of large sets of social media data, Twitter created its "Data Grants" initiative that provides researchers with access to public Tweets to answer their research questions. The grant was awarded to a wide range of projects including, "Foodborne Gastrointestinal Illness Surveillance using Twitter Data," "Disaster Information Analysis System," "The Diffusion and Effectiveness of Cancer Early Detection Campaigns on Twitter," "Do Happy People Take Happy Images? Measuring Happiness of Cities," "Using GeoSocial Intelligence to Model Urban Flooding in Jakarta, Indonesia," and "Exploring the Relationship Between Tweets and Sports Team Performance" (Krikorian, 2014). The variety in current research use of social media data implies an importance in its preservation for historical and longitudinal studies.

Despite social media's apparent value and the existing knowledge on web archiving, there remains much that is unknown preserving social media. Thomson's (2016) report on social media archiving states that, "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 archiving works well for static sites that change infrequently, the dynamic content on social media platforms introduce greater ambiguity (p. 4). Most institutional archives and libraries who use social media for outreach purposes have not tried to preserve it in any way citing the challenges to doing so as a deterrent (Liew, King, and Oliver, 2015). Explorations in the literature provide some initial insight into what issues, challenges and practices currently face social media archives.

1.3.1 Collection Development

Part of any collecting initiative involves deciding what to include in the archive as part of the appraisal process. As we have seen, historically this had led to gaps in the historical records. So naturally this is a difficult task to approach when starting a collection intended to address some of these gaps. In her report on web archiving, Truman (2016) found that "collections based on specific themes or topics require more scoping effort to determine which sites are to be collected" (p. 17). The respondents were frustrated that there was not better communication about what had been collected by other institutions. They explained that a lack of collaboration and communication resulted in "fragmentation of collections or potential duplication of effort." (p. 17). Other institutions did not mind the duplication of effort but would prefer the opportunity to make informed decision. Truman's (2016) report also called for the development of a tool that would make it apparent where archival holdings existed. This study also found that efforts to capture historically significant events often pose too much of a burden for a single institution to adequately capture when events change quickly and websites are at a

greater risk of being taken down (Truman, 2016, p. 19). These curatorial issues could easily apply to social media content specifically as well as to web archiving in general.

Indeed, Thomson's (2016) report on social media archiving remarks that "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" (p. 23). She says that there are a few ways that collections handle this ambiguity. One is by creating a narrow--or restrictive--selection policy. For example limiting the scope to particular accounts which ensures an institutions only collects relevant data. Another way is to create a broad set of selection policies, for example allowing the scope to include a certain location or hashtag which will ensure that more of the conversation is captured. Either way some content might be lost. Selection difficulties for social media face a double edged sword, "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" (p. 24). In addition, these concerns about what content to collect from social media overlaps with some of the legal and ethical concerns described below.

1.3.2 Social media collection tools

To address some the challenges introduced by the dynamicity of social media content, there has been an increase in the development of tools designed for web archiving broadly, and archiving social media platforms specifically. Though by no means intended to be an exhaustive list, some platform specific collecting named by collecting tools named by NCSU's Social Media Archiving Toolkit (2015b) include:

- **Archive-It.** A web archiving tool run by Internet Archive. Its use requires a subscription fee to harvest seed websites assigned by the subscriber. Subscribers can also set a crawl schedule to regularly harvest websites. Archive-It also supports storage and discovery, through their layer known as the Wayback Machine.
- **ArchiveSocial.** Also a subscription service run through the Internet Archive. ArchiveSocial collects original posts and associated comments on Facebook, Instagram, Twitter, LinkedIn, and YouTube. It continually harvests throughout the day so that it does not miss any changing content. It supports preservation and searching, while also allowing users to interact with preserved content.
- **Social Feed Manager**. An open sourced tool created by Dan Chudnov and team at George Washington University harvests Tweets through its public API. It supports harvesting using hashtags, keywords, users or geolocations. Its datasets can be downloaded as CSV files.
- **Twarc.** An open sourced Twitter archiving tool developed by Ed Summers at the Maryland Institute for Technology in the Humanities, Twarc is a command line tool that harvests Tweets as JSON data.
- **Twitter Archiving Google Sheet (TAGS).** This free Google sheet, created by Martin Hawksey, runs through Twitter's API to collect tweets on selected hashtags.
- **Perma.cc**. A free tool that captures and provides a link to a single static page. It is primarily designed as a citation tool to provide a permanent links for researchers that want to cite a website exactly as it existed at the time of their research. It does

not schedule crawls or follow links or documents that are embedded into the site like Archive-It or other web capturing tools. (http://perma.cc).

- **Lentil**. An open-sourced Instagram harvesting tool created by North Carolina State University Libraries and harvests images from Instagram. It also provides an access layer that supports browsing and sharing images. It has a built in moderating and harvesting system for content.

Additionally, some platforms, like Facebook and Twitter, provide their own personal archiving tools to for individuals' data.

1.3.3 Access

Archives traditionally provide access to analog materials on-site in a reading room. Items are made discoverable through finding aids which describe the content and organization of collections. However, social media and other born-digital material can be made widely available online. The content can be linked through the finding aid, or even embedded within it. For web archives, most provide access to content online, while a smaller amount provide access to content in a controlled setting like a reading room (Truman, 2016, p. 24). Most of those that provide access via the reading room are European repositories that have stricter Right to Be Forgotten laws. Web archiving institutions have complained about the discoverability of web archives on several accounts (Truman, 2016). It is difficult to search web collections across several institutions; users may not know what institution houses the records for a specific archive. Additionally, third party vendors usually only allow for URL based searching which means that a user must know what they are looking for before they start searching. This last issue may be less of a problem for social media archives.

Regarding social media content specifically, it is important to provide enough metadata for a social media dataset to preserve its context (Thomson, 2016, p. 24). This metadata should provide information about how a dataset was harvested, when it was harvested, and what the Terms of Service were at the time of its collection. This not only ensures adequate preservation, but also makes searching easier for users. However, the size of social media dataset can actually be an impediment for its use. For example, the Twitter Archive at the Library of Congress (which houses all Tweets), is so large that there has been no way to index the collection to facilitate quick searches (Thomson, 2016, p. 25).

Some tools that have been used to provide access to social media include traditional content management software like ContentDM, online content management software like Omeka, or the Internet Archive's access platform known as the Wayback Machine (Truman, 2016).

1.3.4 Legal and Ethical Concerns

For institutions looking into collecting social media one of the largest hurtles appears to be the legal and ethical issues. A good review of this landscape has been made by North Carolina State University (NCSU) Libraries' Social Media Archiving Toolkit (2015a). Firstly because there is so much of it, it is hard to contact and get responses for permission to archive their data (NDSA, 2014). Secondly, many users sign over rights to their data to the social media platform in their terms of service (NDSA, 2014). This makes it unclear whether or not repositories should seek permission from the social media platform to collect or from the users themselves. However, copyright concerns regarding use of social media data can be largely mitigated, as long as it falls under the

Fair Use (17 U.S. Code §107). Especially, in the case of research analyzing large aggregates of data, copyright concerns are minimal to none (Thomson, 2016, p. 17). The ambiguity remains over Terms of Service which are different for each platform, and change frequently making it necessary to frequently review them to assure compliance. Often though, platforms limit how much data can be harvested at a time through their API. This technical limitation is usually built into the API in order to satisfy the conditions of their Terms and Service (Thomson, 2016, p15-16).

More difficult questions arise when considering the ethics of collecting social media data. Many of these concerns regard users' privacy, and consent regarding their research use in large data sets. If an institution selects social media content using hashtags, for example, then it would be very difficult to seek permission for long-term preservation and use from each user. Summers (2014) discusses a good way to ethically provide access to social media content using an activist archive via a method known as hydration. By going through the Twitter API, researchers can "hydrate" archived Tweet IDs with rest of the data for that Tweet. Hydration reflect users' choices to delete content after the TweetID was archived. This means that any deleted content will also be removed from the archive.

1.3.5 Issues specific to activist groups

In addition to social media specific challenges, institutions initiating an activist-oriented collecting program will face the unique challenges that come from working with activist groups. Springer (2016) describes how the Federal Bureau of Investigation (FBI) preserved records of black feminist group during 1970s, most likely to monitor their activity. With a history of government monitoring such as this, activist groups may be

wary that archiving opens up an opportunity for surveillance. Additionally, activist events may involve acts of disobedience or illegal activity that they may not want known. Carter (2006) explains that organizations may choose to assert power over a dominant group by invoking silence. Filling silences may introduce further harm to a group as "records may no longer have the function or meaning intended by the original record creator (p. 226). As we learned earlier, community archives are first and foremost generated and defined by the community itself to challenge the dominant historical understanding (Caswell, 2014). Protecting privacy and the wishes of individual organizations is important to engender good will, as archives are often situated within institutions that hold social or political power.

1.4 Research Questions

The challenges faced by collecting social media and the records from activist organizations speaks for the need to study the practices, challenges, and solutions used to create and maintain collections of activist social media content. Some specific questions this exploratory study aims to answer include:

- What tools (software or otherwise) are most commonly used to harvest, preserve and provide access to social media content?
- What challenges, if any, are associated with an activist oriented collecting initiative?
- What metadata, if any, are being used to make social media data accessible in the long term?
- What significant challenges did the collections face during the process? These might include challenges such as activists' unwillingness to be documented within

a mainstream institution, difficulty understanding the legal and ethical considerations present, or lack of resources available to support the project.

- How do the collections approach solving identified problems?
- What outcomes did repositories experience as a result of their collecting activist social media content?
- What areas do archivists need more guidance on that could be the focus of future research and professional attention?

Discovering answers to current practices and challenges faced across activist social media archives, will hopefully identify common practices, guide the creation of future social media archives, and help established repositories seek mutually beneficial relationships with activist community groups.

Methods

Since the goal of this study is to conduct exploratory research to identify common practices, tools used, and challenges for activist social media collections, the methods used assessed both what is being done to archive activist social media content and why it is being done that way. Determining the practices used across similar projects indicates commonalities of practice for similar projects harvesting, preserving and providing access to activist social media collections. To accomplish this goal data for the study was collected using surveys and semi-structured interviews. According to Babbie (2010) one strength of survey research is that it allows researchers to describe participants' characteristics. In contrast, the data acquired via semi-structured interviews evaluates why these characteristics occur. This is because one strength of qualitative research is that it captures nuance and allows for depth of understanding (Babbie, 2010). Semi-structured interviews are one way to conduct qualitative research.

2.1 Participants

The participants in this study were archivists or other professionals who work with social media collections from activists groups. To recruit survey participants, I sent an email over SAA Web Archiving Roundtable listserv, which is a fairly active and engaged listserv. I also Tweeted links to my survey, and identified several archivists who had publicized their activist archives and contacted them directly using the recruitment message I drafted for the Web Archiving listserv. To recruit participants for semi-

structured interviews, all survey participants were asked if they were willing to participate in an interview with the researcher. I then contacted those people to schedule an interview who indicated that they were willing to be interviewed.

Thirteen archivists completed the survey, and three archivists participated in semi-structured interviews. Based on the qualitative responses to both data collection methods, at least some if not most of the participants were employed by a large library system usually at a college or university but not always. One respondent stated "we are not a library," implying that they may be from a museum or other similar cultural heritage institution. More specific information about archivists' employing institution was not asked in order to preserve anonymity. So few libraries and archives collect activist social media content, that it might have been possible to match responses to particular institutions, or even particular individuals at those institutions.

2.2 Data Collection

All participants were asked to consent to participate in both the survey and interview portions of this study. The surveys and interviews were loosely based on the survey and interview guide developed by Zach and Perri (2010). Their survey and guide for a semi-structured interview was used to determine the electronic records management practices in colleges and universities. I chose this this study as a model for developing my survey and interview questions for two reasons. Firstly, they explored the practices common in, what was at the time, a relatively new area in archives. My study attempts to do the same. Additionally, they used the same data collection methods (i.e. surveys and semi-structured interviews) that I am using. Truman's (2016) report on web archiving

also served as a model for the type of questions asked during the semi-structured interviews, because of its similar subject area.

Participants first took the 10-20 minute online survey. The survey evaluated whether or not the patterns and practices found in the existing literature hold true across activist social media collections. Questions cover topics such as the size of the collection, what software or web tools have they used to harvest, preserve and provide access to web content, what social media platforms they have collected, their communication style with the activist group they are archiving, what the biggest challenges they faced were, and their funding sources (See Appendix A). These questions placed responses within their context (e.g. type of materials, resources at their disposal) and identified basic procedures for handling social media content within that particular repository.

After completing the survey, I provided participants with an option to participate in 30-60 minute semi-structured interview via telephone. With consenting semi-structured interview participants I asked questions based on concepts and patterns mentioned in the literature review. Questions covered project workflows, ethical considerations, why and how they faced challenges, and any outcomes they have seen from the project (See Appendix B).

2.3 Data Analysis

Data analysis was done by determining the descriptive statistics for responses to questions on the survey. Mostly this involved determining the frequencies of responses on particular question. Frequencies are helpful in determining if certain challenges or practices occurred across many repositories or if they were anomalous.

To conduct data analysis on the semi-structured interview responses I followed Zhang and Wildemuth's (2009) suggested practice for the analysis of qualitative data. They outline eight steps for analyzing qualitative data.

- *Prepare the Data.* To prepare data for analysis I will obtain permission to record interviews from both participant groups using a digital recording device. From the recordings, I will transcribe the interviews.
- *Define the Unit of Analysis.* A unit of analysis refers to the "basic unit of text to be classified during content analysis" (Zhang & Wildemuth, 2009, p. 310). For the interview results, I used a single theme identified in the transcripts.
- *Develop Categories and a Coding Scheme*. I used inductive reasoning to develop categories and code for the data because this is exploratory research rather than proving a hypothesis. These categories matched areas of interest as well as common themes that developed across participants' data. These themes are included in the results section of the study.
- *Test Coding Scheme on a Sample of Text*. To determine the appropriateness of the coding scheme I use, I tested it on a sample of my data to ensure its effectiveness. This was an iterative process.
- *Code all the text.* I will use the coding scheme developed in the previous step to code all of the interview transcriptions I will collect.
- *Assess coding consistency.* Though I did not have to worry about interrater reliability in this case, I wanted to account for my own bias and the change in my understanding and application of the coding scheme over time. To do this I reviewed what I coded and made changes to ensure consistency.

- *Draw conclusions from coded data.* Coding the data in the previous step allowed me to analyze descriptive statistics as described for the survey data collected in this study. It helped assess what practices have been successful across the repositories, what strategies they used to overcome those challenges, as well as their assessment of the outcomes for their projects.
- *Report methods and findings,* which I have done in this section and in the results and discussion section of this paper.

Survey Results

Participants responded to survey questions that assessed nine different variables related to their institutions collecting initiative that resulted in the acquisition of activist social media content. Participants were given a chance to clarify their responses to each question. The following are the descriptive statistics for each variable integrated with their qualitative responses to provide further clarity.

3.1 Digital Projects

Eleven out of eleven respondents already had digital collections programs at their employing institution. In many cases, these digital collections were sizable. Based on their open-ended responses when asked to describe the digital collections at their institutions, the size of respondents digital collections range from a few thousand digital files (e.g. “I believe we have 10K items in ContentDM”) to very large (“The web archiving portion is around 750TB”). Since these archivists already had digital collections at their archives, it implies that there is already some process in places for ingesting, preserving, and providing access to digital content. This may have alleviated some of the barriers to starting an activist collection project with a social media component for the respondents.

3.2 Process and Workflows

Many survey questions focused on how the participating archivists collect, preserve, and provide access to social media. By knowing how other institutions develop social media collecting programs, it may lower the barrier for other collecting institutions

that want to start similar programs but are daunted by perceived complexity or technological difficulty.

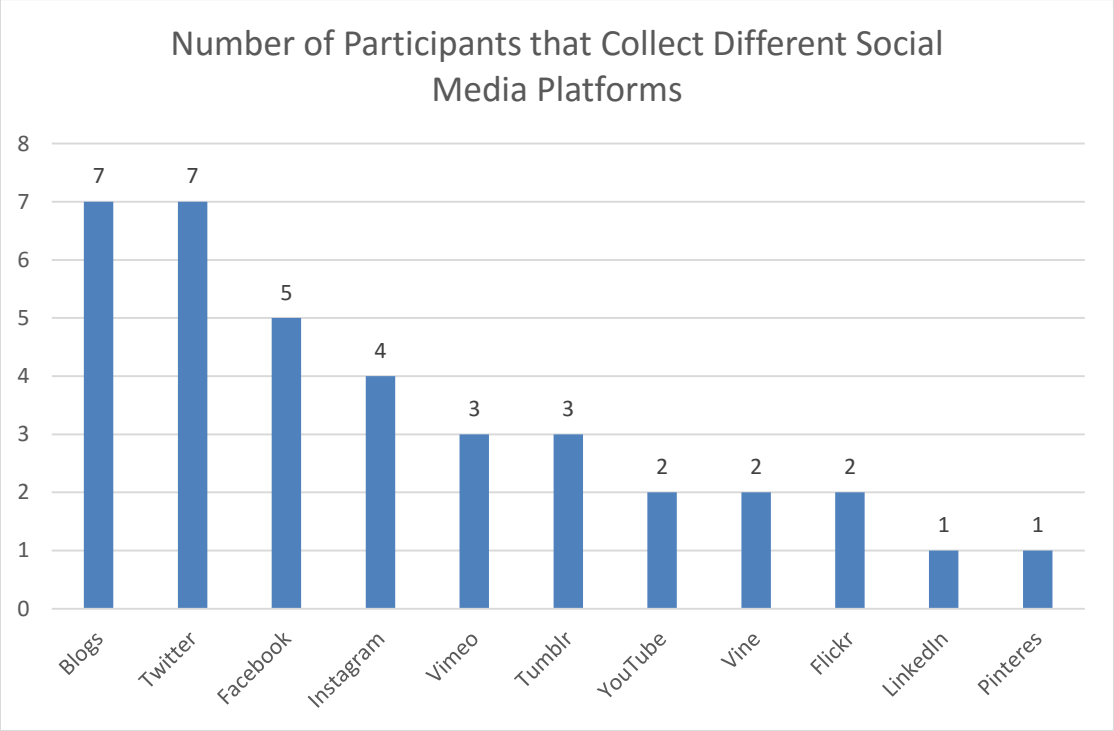

FIGURE 1: Number of participants that collect different social media platforms.

To determine what social media platforms are currently being collected by archivists, I asked respondents which platforms they collect for their project. Twitter and blogs sites are the two most commonly collected social media. Of the ten responses to this question, seven archivists collect Twitter (70%), seven archivists collect blogs (70%), five archivists collect Facebook (50%), four archivists collect Instagram (40%), three archivists collect Tumblr (30%), three archivists collect Vimeo (30%), two archivists collect Vine (20%), two archivists collect YouTube (20%), two archivists collect Flickr (20%), one archivist collects Pinterest (10%), and one archivist collects LinkedIn (10%).

Two participants chose to explain their answers in the space left to specify other social media responses. Both explained that they were technologically unable to capture

"Twitter and Facebook feeds" because they do not have an installation of any web archiving software. Currently, both participants only capture blogs. This implies that blogs are the easiest social media platform for respondents who did not have access to more versatile tools, such as Archive-It.

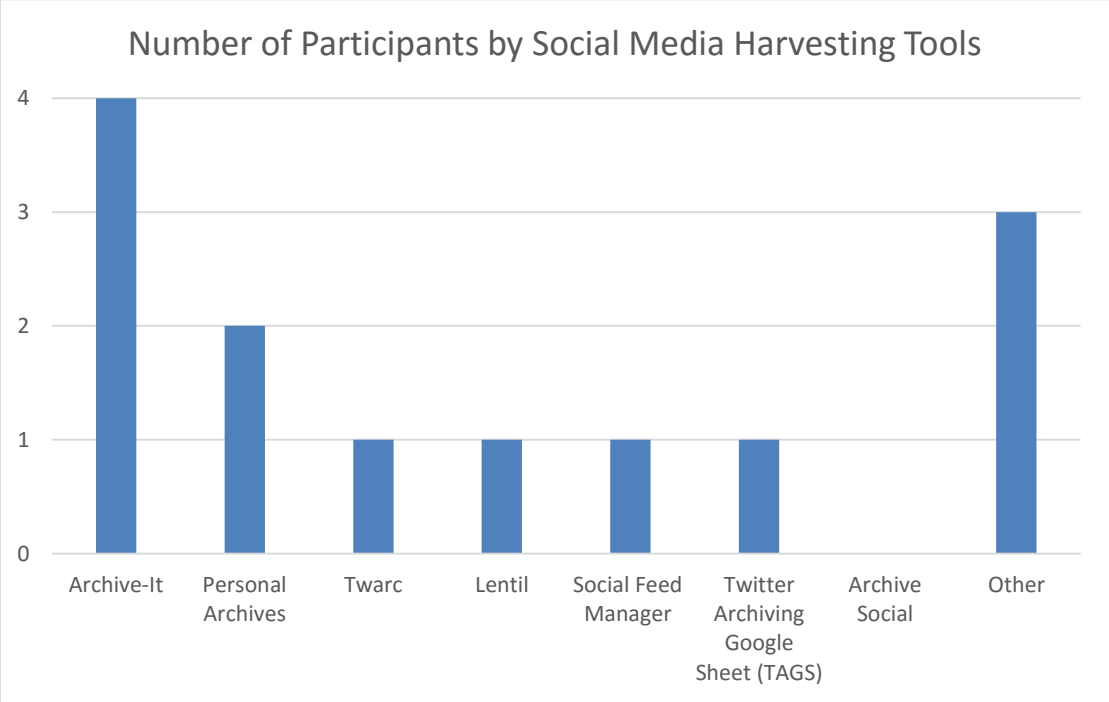

FIGURE 2: Number of participants by Social Media Harvesting Tools

Of the tools available for harvesting web content, Archive-It was the most commonly used among participants for capturing social media. Of the eight archivists that answered this question, four responded that they use Archive-It (50%), one uses Twarc (13%), one uses Lentil (13%), one uses Social Feed Manager (13%), one uses Twitter Arching Google Sheet (TAGS) (13%), and two use personal archives that have been downloaded by the user and donated (25%). No respondents use ArchiveSocial (0%). The second most frequently chosen by archivists was 'other' (38%). When asked to explain what other tools they use to harvest social media, one archivist (13%) said that

they "Use the Internet Archive as a contractor for web archiving, which is much like Archive-It but not the same thing." The other two archivists responded that they use perma.cc (25%).

The metadata used for the social media content they harvest reveals more about how each institution preserves and describes the social media content they collect. Of the eight archivists that answered the question about what metadata standards they apply to their activist social media collections, four responded that they do not use any metadata standard. Three of these four did not specify what if any metadata they collect. However, one explained that their harvesting tool (perma.cc) does not provide the option to provide additional metadata other than what it automatically generates.

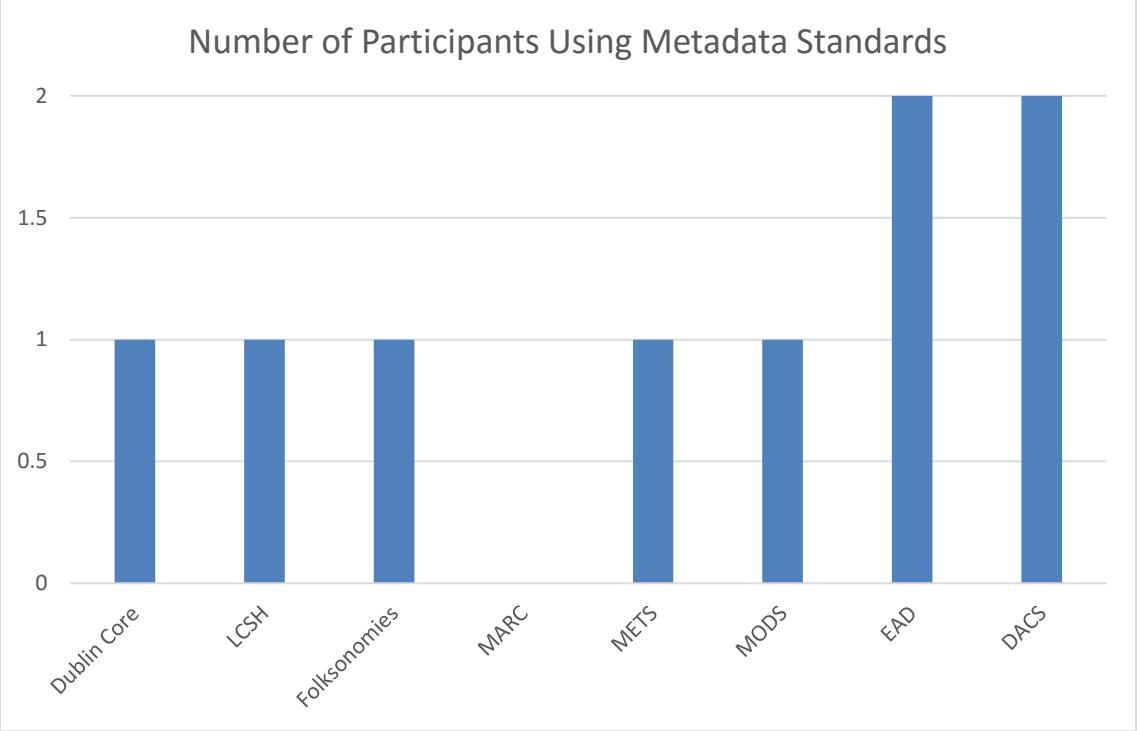

FIGURE 3: Number of participants using metadata standards

None of the archivist respondents reported using the MARC to describe their data. One archivist reported using Dublin Core (13%). One reported using a combination

Library of Congress Subject Headings, Metadata Encoding and Transmission Standard (METS), and Metadata Object Description Standard (MODS) (13%). Another archivist reported using a folksonomy such as user generated tags (13%), while two archivists reported using both Describing Archives a Content Standard (DACS) as a descriptive standard and Encoded Archival Description (EAD) as the encoding standard for the finding aids where the social media content is described (25%). Overall, the lack of consistency implies that there is not one specific standard best used for social media data. An institution with a social media collecting program would not have to adopt a new metadata standard to fit it into their existing digital preservation structures.

By far the most common way archivist provide access to activist social media is through archival finding aids. Four of the eight archivists reported using finding aids (50%). All four of them provided more information about how access would occur in the open response section available after this after this question. One of these revealed that they have not begun providing access to social media content, but that it will be discoverable in a finding aid, and access will likely occur onsite in a reading room. Another of these respondents revealed that they do not have a way of providing access to born-digital material through the finding aid, but they are able to for digitized collections at their repository. The way she/he described it is, “We provide access to digital imager and some documents through finding aids, but we do not have method for serving up born digital material in any meaningful way, yet (especially not social media content).” Two of these respondents explained that they have another discoverability tool in addition to finding aids. One uses the Wayback Machine, which is the publically searchable Archive-It interface, and the other uses the public link provided by perma.cc.

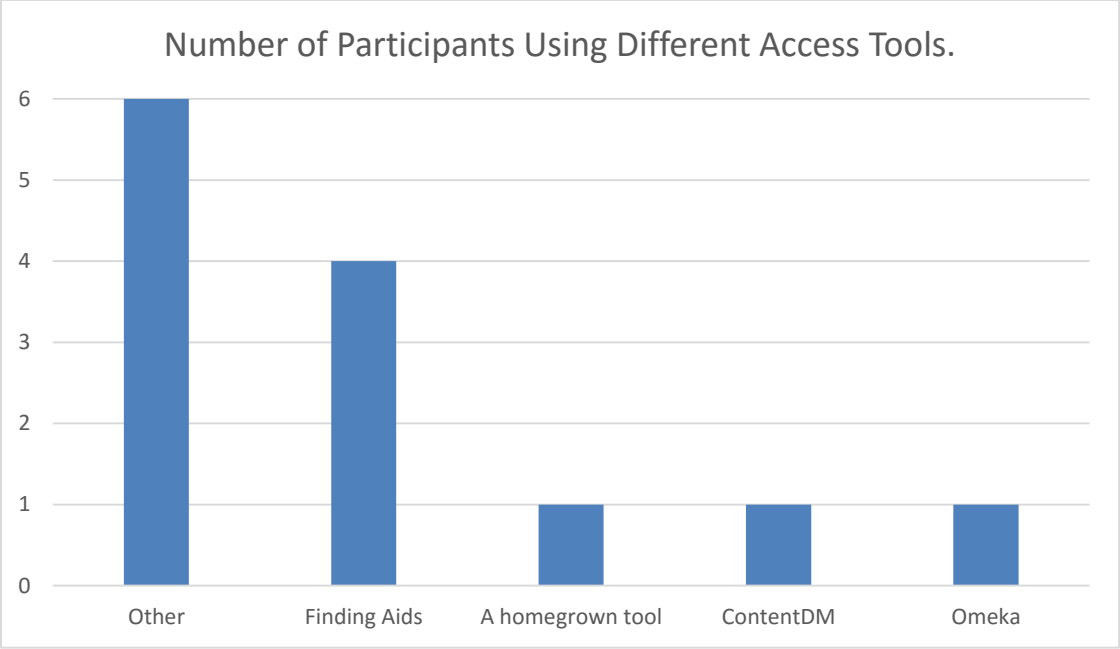

FIGURE 4: Number of Participants Using Different Access Tools

Other methods for providing access included one respondent who uses Omeka (13%), one respondent who uses ContentDM (13%), one respondent who uses a homegrown tool (13%), and one respondent who described using another tool not provided as an option on the survey (13%). This last respondent clarified by explaining that "perma.cc is public," perma.cc being their harvesting tool.

3.3 Project Management

Adding a new collecting initiative at an existing archive or cultural heritage institution would naturally require aspects of project management, like managing staff time and available funds. Questions regarding project management reveal how much, if any, extra staff or time social media collecting activity accrued for these projects. In the surveys participants were only asked about the source of their funds. More information regarding project management issues came out during the interviews.

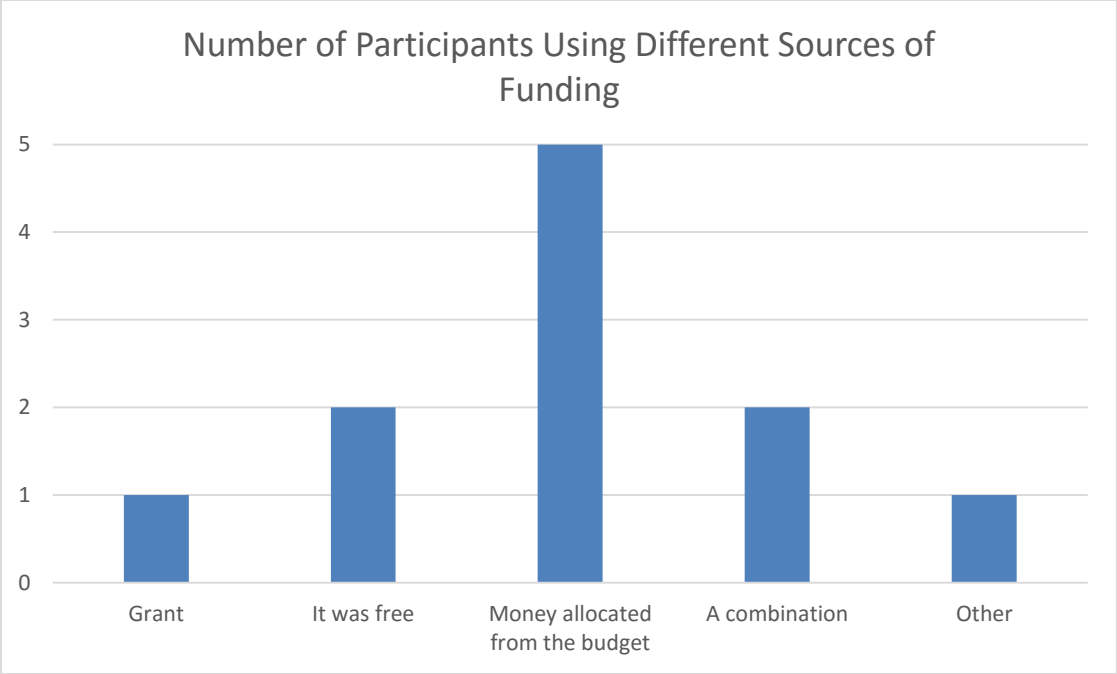

FIGURE 5: Number of Participants Using Different Sources of Funding

Eleven participants answered this question. The most common source for the projects funding was "Money allocated from the library or department budget." Five archivists chose this answer (45%). Presumably, this response could mean that it fell within the normal work activity supported by their unit, or that more money was provided by the department for the project, as one participant clarified the project was "part of the web archiving service already in place for the library." Two archivists reported that it was free (18%). Perhaps indicating that it fell within their normal work responsibilities and/or used free tools. One respondent had the work supported by a grant (9%). The one respondent who marked "other" explained that he/she "shifted priorities to focus on this work in my day to day role." This is similar to how respondents interpreted the option: "Money allocated from the library or department budget." Finally, two respondents indicated that it was supported using a combination of these methods (18%).

3.4 Challenges

Eight participants chose from a list of possible broad-challenges that could occur when collecting activist social media. The most commonly reported challenge was concerns over Legal Issues. Five participants claimed that this had been a challenge for their project (63%). Four participants reported that ethical issues had been a concern (50%), while three people reported that harvesting data was a challenge (38%). Finally, only two people found funding their project difficult (25%).

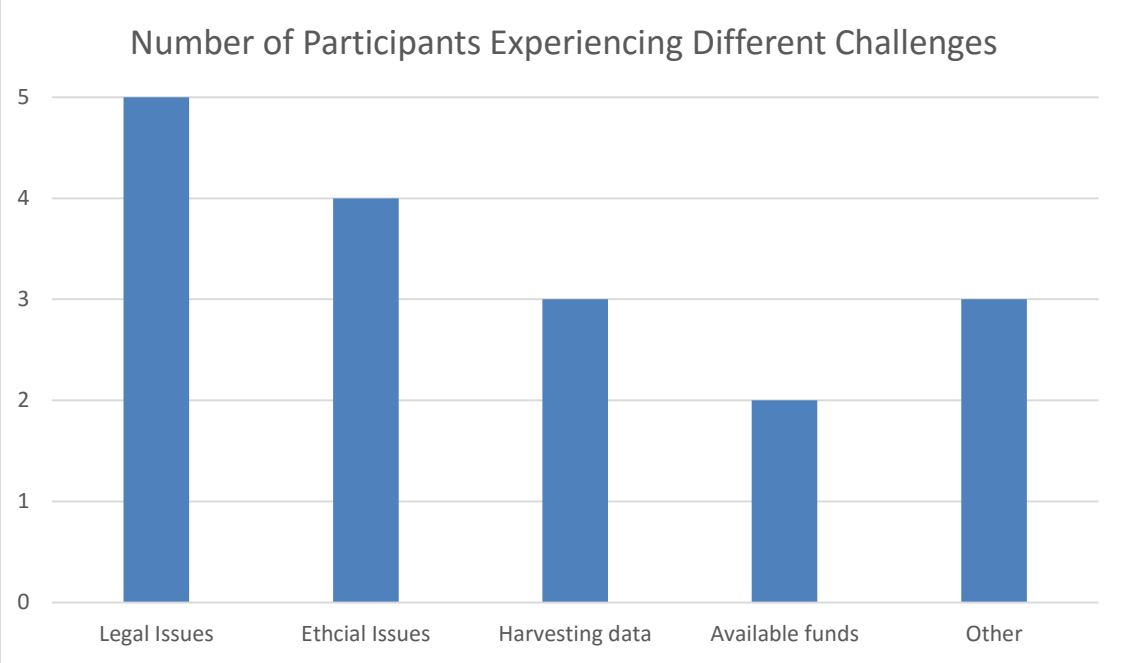

FIGURE 6: Number of Participants Experiencing Different Challenges

Three people claimed other challenges, which they elaborated on. Two of these participants described issues that might be described as "curatorial." For example, one respondent explains that for them it is difficult to determine the value of different platforms taking into consideration how different activist groups use them. She/he explains that for sites like Twitter of Facebook sometimes it is enough to capture a small sample of what is there; accurately capturing the whole page does not add more value.

Whereas, for him/her, “a separate problem exists with YouTube and Vimeo, which are usually lumped in as ‘social media’ but can be carriers for serious content not available elsewhere.” For social media related to a particular activist organization that they capture “it can be a problem to identify them in the first place…Again, the problem with this is mostly knowing it is there to go after.” Similarly another respondent explicitly describes their challenges as “curatorial, like time and awareness to develop a comprehensive profile of accounts to be tracked.”

The third person described their main challenge being that their institution is expected to adopt Archive-It but has taken a long time to do it. As a result the current tool they use is good at capturing simple pages but dynamic pages do not work as well. So they are having trouble capturing a wide variety of pages because they are waiting for an institution-wide tool adoption.

3.5 Outcomes

Only five participants responded to the question about project outcomes. Some participants indicated in the free response section that their project was still in the early stages and they did not have outcomes to report. This may account for the lower response rate for this question.

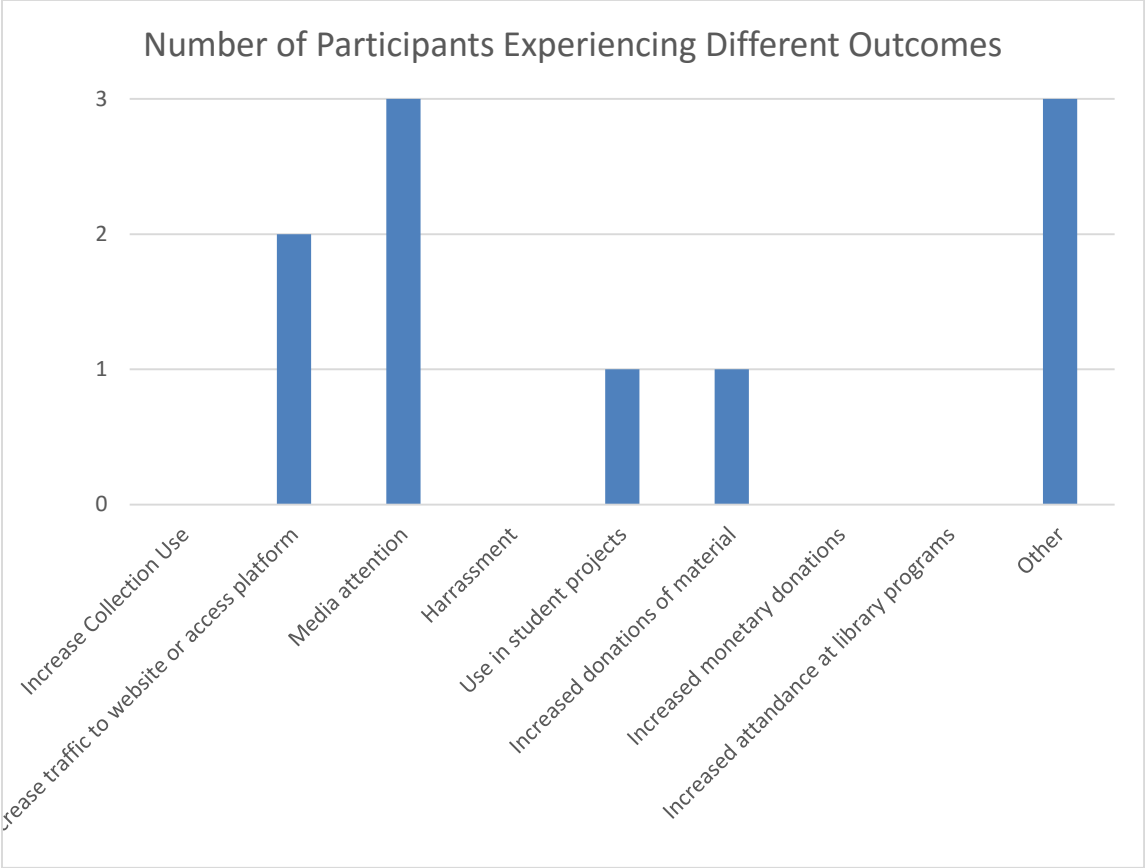

FIGURE 7: Number of Participants Experiencing Different Outcomes

Three respondents indicated that they had received media attention regarding their activist oriented collecting initiative (60%). Two respondents cited increased traffic to their website or access platform as a result (40%). One respondent indicated that the collection was used in a student project (20%), and one respondent claimed an increase in donations of material to their repository (20%).

Three respondents named other as one outcome of their activist social media collection. In the space provided to clarify this response, one participant explained that before they had no web archives so in that way usage has increased, “but usage is moderate at this point.” Another participant explained that there has been “increased student awareness of the archives and its mission. One student even offered to help

promote the archives even more to the student body." The last participant that marked other explained that they have not promoted the project so they have not experienced any outcome yet. Two other participant who did not check the "other" option also explained a lack of outcomes because the project had not been promoted yet.

3.6 Areas of ambiguity

The participants responded to a free response question asking them what aspects of social media collecting needs to be examined further by professionals and the professional literature.

TABLE 1: Areas of future research identified by Participants.
Mainly of the tools are evolving in use and value.
I myself am curious about what the use cases in the future for Facebook pages will be. Is it going to be drilling down and seeing the exchanges of annoyance and anger about social issue X on the Facebook of some organization? Or use of a large set of Facebook pages for text analysis? Or evidentiary, that is, to demonstrate that despite statements to the contrary, organization X did say Y on date Z. For a particular organization, I can see that having an accurate archive of the organization's own Facebook etc can be useful if not critical for evidentiary and records management type reasons. However if we are talking about large publishing organizations, I am not so sure (as I have already suggested) that worrying about social media isn't a distant second to the need to archive the organization's site fully and accurately, which I think is assumed to be simpler than it often is for organizations with big sites.
Data preservation and privacy issues are always tough
More attention needs to be invested into obtaining consent of content creators.
We'd like to perform a more thorough analysis of the ethical considerations of harvesting and preserving social media data, especially of social media users who are not aware of our collecting initiative. Generally speaking, the archival community could develop some "better practices" when it comes to collecting, describing, and providing access to these sorts of data. How institutions deal with copyright & privacy for harvested social media varies quite a bit. I'd like to see some more defined best practices for this.

As Table 1 indicates, only one participant was curious about the usability and value in different harvesting tools. The most mentioned area of future study was the ethics of collecting social media, particularly activist social media. Whereas, the most common area where people would like to see more research was on ethical issues surrounding consent and privacy for the people or groups whose social media they collect. Two participants called for defined best or better practices surrounding this issue.

Interview results

I conducted three interviews with archivists collecting social media from activist groups. Some commonalities exist between the settings and the projects, while there are some differences as well. Overall, the experiences of these archivists provide some insight into challenges, concerns, and continued areas of ambiguity for archivists collecting activist social media. The similarities in the way they have imagined their projects may provide some guidance for similar projects at other places.

All three interviewees are digital archivists at universities with large library systems. Each project is referred to by an assigned project number. I have done my best leave out information that might identify the specific project, or the people who work on the project. I will first describe the major contextual aspects of their activist collecting initiative or social media project, followed by a discussion of each projects responses on seven themes that developed. The main themes I identified were 1) collection development, 2) access, 3) project management, 4) acquiring permission or consent, 5) challenges, 6) outcomes, and 7) areas for future research.

4.1 Project 1

The institution where Project 1 is held has been collecting social media for several years. The university library collected Instagram photos related to student engagement with the library. After completing this project, they applied for and received a grant to explore social media collecting further. During work on the grant, a violent event occurred that affected the campus and the local community. They collected select Twitter

and Instagram data from a specific hashtag related to this event. Because of the hatred involved with the event, there were many posts under these hashtags that related to activist or advocacy messages documenting both sides of the social conflict. So unlike other two projects, Project 1 did not start as a specific activist-oriented collecting initiative, and is not focused on a particular group.

4.2 Project 2

At the university where Project 2 is hosted, there was increase in activism on campus during 2014 and 2015. This was not the only university to experience an increase in activism that year. One or two student groups held sit-ins on campus in the Fall of 2015. Frequent turnover in leadership for these student groups due to graduation, meant that it was important to collect their records soon. The university archive launched an initiative to collect the records of activist student organizations at the University. The archives held two collection drives, one in the student center and one in the archives which is farther away from the center of campus and would offer group members some anonymity. When they advertised these drives they explained that they would take any kind of record, from traditional analog material to email archives and social media accounts. They had a better turn out at the drive held in archive, which surprised them because much fewer students visit the building or know much about the archives.

4.3 Project 3

The archive hosting Project 3 is affiliated with a particular school at the university. They also saw an increase in activist activity during the fall of 2015. They were concerned about losing valuable information on the student body, so they began an initiative to collect student papers, including social media content. Before they started

this initiative though they began collecting the blogs of these student groups, they have only recently began outreach efforts publicizing their activist student group collecting initiative and have not received any new materials yet.

4.4 Collection Development

Project 1. The university archives at this institution collects social media using a variety of methods including Social Feed Manager, Twarc, and a home-grown software tool. Social Feed Manager collects social media accounts, which they use to collect the accounts of different University affiliated accounts. However, most of their activist content comes through their hashtag based collecting. The other two tools also generate a dataset.

Project 1 was also very clear about how they decide what social media content to collect:

> The way that we've tried to approach our collecting is to align decisions about what data to harvest, to align those with established collecting strengths… [we] build collections in those areas, and the accounts, and keywords, and hashtags should ideally map to one or more of those existing collecting areas… we're not collecting all of Twitter which was not the work we wanted to do.. So how do you cut into this social media? Well, what do your researchers want to see? What your researchers might be expecting is just the content you're already collecting.

Project 2. The archivist on this project said, "I could see a case where someone had a set of records that they would feel more comfortable discussing and donating in a more private space so I'm glad we stuck to having this open as an option, and I think if we do another collecting drive on this initiative or another sensitive topic I think we would likely forego doing it in the student center." Once they gain permission from their donors, in this case from activist student groups, they crawl different social media

accounts using Archive-It. This specific project was the first time that they had crawled a Facebook page.

Project 2 did also identify some relevant social media content to be included in their collection. They crawled relevant social media feeds from on campus news organizations that provided coverage of activist events, especially during two very active days in Fall 2015, however they hid the crawled content from public view. Afterwards they sought permission from these organizations to preserve and provide access what they harvested.

Project 3. When they were deciding what content to harvest they focused on student blogs. They said, "Student blogs have been a really big part of the longer form conversation around what's going on." This project was one way that Martin imagined trying to diversify the collection as they have no other material related to student activism in their school's library. An activist oriented collection diversifies the collection, "in the sense of who racially is represented but also by the types of materials we're pulling together." Their archive has always said that it would collect student papers, but has not done so up until now. Their archive uses perma.cc because it is a free tool that easily and accurately captures web content. They are hoping to begin working with Archive-It soon, once their whole library system adopts its use.

4.5 Access

Project 1. Though this institution has collected and preserved the data from hashtags they have not yet opened what they have harvested to the public. They have been struggling to imagine an access method that feels comfortable. One that covers concerns regarding consent from all users whose social media posts are included in the

collection. They know that it would be difficult and time consuming to get express permission from every social media user who posted using the hashtag they are harvesting. For this reason, they do not feel comfortable sharing the data on the open web. Additionally, there are social media Terms of Service conditions they might be violating providing that much content online. They are imagining moderated access where, though the content might be publically available, a researcher would only be able to view the data in their reading room, or by providing Tweet IDs that the researcher could then "hydrate" themselves through the Twitter API. The archivist on Project 1 says, "I don't think we would make the Instagram content available on the open web except for cases where we got permission from the people who either took or posted the photos." The data that they collect would be discoverable through the university's EAD and DACS compliant finding aids once they do begin providing access.

Project 2. To provide access to the collections Project 2's archive is using Finding Aids and the Internet Archive's Wayback Machine. That means that people can discover the social media content through the finding aids at the University or through keyword searches on the Wayback Machine. They have mirrored the metadata about each collection on both the Finding Aids (using EAD and DACS) and the Wayback Machine (using Dublin Core). Four collections from different activist student groups have already been made available at his University.

Project 3. Similarly, the archive at this university plans to provide access through their archive's finding aids (using EAD and DACS), which are also found in their library's OPAC. They can embed the links to the archived perma.cc page on the finding aid. Their archive will not provide access to the social media content until they talk to the

student organization and receive any other material from them. They say that they are, "open to trying a lot of different access points," but has had a chance to give it much thought so he was not sure what form it might take for social media.

4.6 Project Management

Project 1. The project was mostly funded via the grant that they received with the rest of the cost being absorbed by existing library funds. There have been about five staff members and number of students working on the project. The staff includes three digital project staff from different departments, the University Archivist who is the curator for the collection, and a legal expert. These last two people are only involved on areas of the project that match their expertise when needed. Two of the digital project staff dedicate the most time to the project with support from the third staff member and the other students. The staff members mostly work on developing the software to harvest social media data, since they are using homegrown open-sourced methods to collect the data.

Project 2. There have been a total of five staff members supporting this initiative across the rare books and special collections library at this university. Four of them work in the University Archives and one other person from a different department who helped reach out to student groups to publicize the collection drives. They also have a student who recently started to help out with the outreach efforts among student groups. They consider the work to be part of their normal responsibilities and they have not received additional monetary support for the work.

Project 3. There are two main staff members working on the Project 3 at the moment: the digital archivist, and the curator of the collection that the activist collections will belong to. Most of their efforts so far have been on advertising the initiative, and on

beginning a web archiving program since they could not wait for Archive-It to be implemented at their institution to capture this content. They also have the support of the outreach working group at their archive, which is led by the outreach librarian and made up of people throughout the library. They spread news of new projects or programs over social media and other outlets. They also have one student worker helping them harvest blog content. There has been no additional funds given to their department to support this work. They have absorbed it as part of their normal responsibilities.

4.7 Acquiring permission or consent

Project 1. The staff on this project know that it would be difficult and time consuming to get express permission from every social media user who posted using the hashtag they are harvesting. For this reason, they do not feel comfortable sharing the data on the open web. Additionally, there are social media Terms of Service conditions they might be violating providing access to large datasets online. They are imagining moderated access, where though the content might be publically available, a researcher would only be able to view the data in their reading room, or by providing Tweet IDs that the researcher could then "hydrate" themselves through the Twitter API. They say "I don't think we would make the Instagram content available on the open web except for cases where we got permission from the people who either took or posted the photos."

Because the project did not start as a specific initiative to collect activist records and because they are harvesting hashtags, they have not so far worked closely with the advocacy groups on either side of the conflict they are documenting.

Project 2. Staff on this project "didn't want to capture without their consent," but knew that some information would be lost if they did not act immediately so they ran test

crawls using Archive-It and hid the data from public view until they gained consent of the group who generated the data. They said that they chose to follow the student organizations' accounts as opposed to, "following a hashtag. We didn't want a situation where we had a number of Tweets that were not intended necessarily to end up in the archives. So we decided to go down a traditional provenance based approach and directly inquire to student organization about permission to capture their site and make it available. And all of them agreed." In this way, the project model used at their University closely matches the traditional provenance based approach to the acquisitions process.

Project 3. The school archive has just begun publicizing the initiative so has not received any papers from the student organization, nor have they been contacted yet by any organization. They plan to meet with the leaders of the organizations and explain the privacy concerns with them. They said, "we will have a form they will sign. And what we've decided is that we will ask the leader of each student organization that's responsible for the [social] media that we've captured to discuss among their group and then sign off on behalf of the entire group."

4.8 Challenges

Project 1. Time is a significant challenge for Project 1. While the grant was in progress it had been a priority because they had to report back to the granting body on their results. Now that the grant period has ended, "there's nothing externally making it a top priority over some of the other things that we have to build that are also top priorities." For this institution it is hard to find time to dedicate to collecting social media when all of the staff that work on the project have many other responsibilities.

Another challenge that they have faced is ambiguity surrounding the ethics of providing access to the data they have harvested so far. They believe there may be some more to learn about the specifics for different social media platforms' Terms of Service as far as what they can provide access to but also that there is little precedent for how to ethically use social media in research. When talking about archivists collecting social media throughout the profession they say, "We're not an insensitive group of people who are going to disregard ethics, but I think that there's a pretty good momentum in the favor there being a way for it to be done ethically. We don't have all of the answers of what that's going to look like because in a way it's a discussion that's happening with today's researchers who are accessing and using Twitter data." They clarified that some of the use ethics for social media content falls into the professional realm of researchers as much as of archivists'.

They also find certain curatorial aspects of collecting social media to be challenging, particularly when trying to identify what they should be harvesting online. They say "early on when we were applying for this grant, we were told that we're not collecting all of Twitter," and that they try to "align decisions about what data to harvest, with our established collecting strengths."

Project 2. The main challenge for Project 2 matched that of Project 1; it was difficult to find enough time to work on the initiative among other responsibilities. Other than this, they said that they were particularly aware of the ethical considerations surrounding their position of power within the University. They said, "we were aware of the power dynamics at play and we're aware that we're dealing with undergraduate students who on top of leading different initiatives through their organization they're full

time students." Though they are acutely aware of the ethical and legal implications of collecting social media from activists, they feel that they take special care to protect the intentions and identities of the student groups who donated material. But they say that it was largely the same approach they take for all other records, which involves, "getting consent, informing them about our procedure for how we take things in, that we describe them in a finding aid, and when they are available they will be open to the entire world, and letting them know that that's what they're signing up for."

Project 3. Their library is part of an extremely large library system at their university with many decentralized units. The library system has been trying to implement Archive-It for a while, but making sure that their agreement with the Internet Archive covers the needs of all members of the library system has taken a while. They hope to be able to capture more dynamic social media content like Facebook and Twitter once they begin using Archive-It. They cite the size of the library system the primary challenge Project 3 faces. "It's how quickly can you get support from the top to roll out the technology that you need to capture today. I think the biggest challenge is that we can't wait for someone to pass away in order to get their collections these days because everything in the digital world has such a short lifespan." These types of projects at large institutions may be more difficult to implement because of the organization's size.

Like Project 2, Project 3 was concerned with their relationship with the activists as well. They expect that it may be especially important to build relationships with the activist student groups because the archive is situated within the university system. They expect to approach the students groups more actively, instead of passively waiting for people to deposit their materials. "Outreach really is a big part of building the

relationships and making sure that they feel comfortable depositing their records with us." They expressed the importance in communicating to activists that the archive wants to preserve the historical record and not perform surveillance for the university. They plan to host programs with student organizations to explain what archives do and why the content of archives underrepresents certain groups, "We obviously can't keep everything, so whatever we do decide to keep, that's a political act because it's a level of interpretations. It's not the same as what a historian does but there's still an interpretive act when you decide to collect something or not to collect something." They want to make it a "participatory conversation" to help build the relationship between these groups and the archive.

4.9 Outcomes

Project 1. Project 1 has not publicized their social media content because it is not yet public. Therefore, they did not have any outcomes regarding use of the collections to report. However, the main thing they hope results from their collection of social media was that their work might serve as a model for other institutions--that they might "lay out some of the foundational work." Already some of the work related to their grant project has been cited in the literature on collecting of social media. Additionally, they hope that they might model for other institutions how to make collecting decisions for social media content. They say, "that's one way this project could potentially serve as a model for others. How do you cut into this social media? ... Collecting things that extend--make a little bit more comprehensive--the existing collecting areas."

Project 2. Project 2 experiences two main outcomes to report from their collecting initiative with activist student organization. Firstly, that there had been some

student interest in the collecting initiative. One student in particular will be working with them to help perform outreach and publicize the collection. He said that the initiative has "raised questions in students' minds about what records we already had collected pertaining to students' organizations." Secondly, they have received donations of student organizations' records from alumni who were activists during their time at the University.

Project 3. Because they had only started promoting their collecting initiative recently, Project 3 had no main outcomes yet to report for their project.

4.10 Areas of future research for activist social media

Project 1. When asked what about collecting activist social media they would like to see studied further, they explained that as a profession one of the challenges is to establish best practices around ethically collecting and providing access to social media. They would like to see what other institutions are doing and, "presumably we would all learn from each other and something might come of that as best practices."

Project 2. When asked how what aspects activist social media collections need to be studied further he explained that there needs to be documented methodologies to achieve consent and ethical access. They would also like to see case studies of how those methodologies have been implemented. Finally, they explained how it feels disingenuous to only collect social media while it is being covered in the news. They wonder if, "Archivists [who] follow these significant events and try to document them immediately, what are the advantages and disadvantages to that." They want to know if "jumping in and doing something is creating more harm, is creating more silences, creating more gaps, creating more vulnerabilities for people to be surveilled, to be harassed." They

think that it may be more beneficial to develop long-term partnerships with groups before their activities become “trendy.”

Project 3. When asked what areas need to be studied more when it comes to activist social media, Project 3 wanted to learn more about how other people approached getting consent and privacy. They said at one point they thought they “might have to reach out to every person whose account we capture, on say, a Twitter feed.”

Discussion

The goal of this research study was to determine common practices used by the professionals that collect, preserve, and provide access to activist social media content. The similarities and differences across the participants identify areas for future research and suggest the development of best practices. The primary findings reveal that activist social media collecting projects: 1) face ethical and collection development challenges, 2) usually follow traditional models for acquisition, description, and access, and 3) increase donor and user engagement with collections.

5.1 Challenges in Ethics and Collection Development

The challenges identified in the survey responses differed from those identified in the interview data. Specifically, the survey responses indicated that collection development was a significant challenge while two of the three projects cited time management as a significant challenge. Yet despite their overt responses, the topics that came up most often across all three interviews were collection development concerns surrounding the ethics of curating social media content and informed consent from activist groups. Most of the survey and interview respondents played a balancing act between capturing data quickly to guard against its loss and receiving permission to preserve the content (e.g "We're definitely going to be seeking consent, but we were just concerned about preservation first and access later…That was a concern about privacy, and if they say no we're going to delete it anyway"; Project 3). The legal issues seemed not to concern most of the respondents, though they admitted some ambiguity

surrounding platforms' Terms of Service: "[Harvesting social media] is a fairly conclusive fair use argument. It's a pretty legal thing to do. I think that there will be some issues about how to interpret certain Terms of Service, but doing it ethically [is more unclear]" (Project 1).

Some archivists discussed activist oriented collecting as a method of filling in archival silences. Both Project 2 and Project 3 explained that their institutions should always have been collecting student records from activist groups, but that they had to start collecting initiatives in order to fill in this gap in their collection. However, whether or not these specific types of projects accomplish redressing archival silences is an important area where future study could occur. Some archivists feel that focusing on activist groups is not enough to redress past collection gaps. For instance, it remains unclear whether creating an activist oriented collecting initiative "is creating more silences, creating more gaps, creating more vulnerabilities for people to be surveilled, to be harassed" (Project 2).

However, most respondents explained that they took extra care surrounding consent and public access because of the sensitive nature of activist collections. For example, Project 2 hosted one of their collection drives in a less public space to preserve anonymity, and all interview participants would only make data publically available online once the groups had consented. The staff on Project 1 summarize the difficulty surrounding collection development, "Getting the stuff is relatively easy. But all of the rules and policies and protocols that support it, those are a little bit trickier."

I can infer from the results that time management and organizational structures might be significant challenges for all archival initiatives but the collection development

and ethical concerns are specific to activist social media archiving. This finding matches conclusions reached by other reports on social media and web archiving which found that the social media platforms' are not aware of how researchers use their social media content (Thomson, 2016, p. 20), and that archivists have difficulty identifying web content to include in topical or thematic collections (Truman, 2016, p.17).

5.2 Practices for Harvesting, Describing, and Access

It might have been expected that activist social media collections would have reported more difficulty in acquiring social media data. However, interview and survey responses indicate that data collection is not the most challenging activity for activist social media archives. This supports findings reported in Truman's (2016) web archiving environmental scan, which indicated that, "tools seem serve some areas of functionality very well – such as capture and analysis. However many of these capture and analysis tools are very specific to narrow types of media (e.g. capturing tweets) or support for particular types of analysis (e.g. link analysis)" (p. 27). As far as the tools and technical skills required to actually collect social media, survey and interview respondents in this study most frequently used Archive-It. Indeed, Project 1 was the outlier in that they sought to harvest large datasets of social media content related to hashtags or keywords in ways similar to the methods described by Thomson (2016) in her report on social media data collection. Both Project 2 and Project 3 harvest accounts using web crawling tools (Archive-It and Perma.cc respectively). Even Project 3 plans on switching to Archive-It in the near future.

Archive-It is a valuable tool for archives because serves many important functions of web archiving: it is fairly easy to use, provides extensive documentation of its

functions, it crawls, preserves, and provides access to harvested webpages, allows researchers to use the Wayback machine to access archived pages across multiple repositories, and it allows a repository to scope crawls to capture as much or as little of webpage as is desired (Truman, 2016, p.40-41). All of this comes with the support of the Internet Archive if any technical difficulties arise. Its primary drawback is that it is a paid for service that could be beyond the means of smaller collecting institutions.

Despite the many positive traits of Archive-It service that attracts many archives to its use, some professionals have expressed concern about relying on it too heavily. Some identified risks that come with outsourcing of web archiving functions, particularly in relying on one service provider such as Archive-It: "We need to be careful about resting on one architecture…multiple methods can deliver differing results…We need to be periodically re-evolving the criteria and methods we use until we can be certain we are the results we need" (Steve Knight, cited in Truman, 2016, p. 41).

It remains unclear whether relying on outsourced services like Archive-It is a significant risk. It could warrant future study, especially when it comes to archiving social media. Certain kinds of research favors the use of datasets over qualitatively analyzing the static content of a web page from a particular time period. Relying too heavily on Archive-It may ultimately result in loss for those researchers.

The survey assessed descriptive practices by inquiring about metadata used to describe activist social media collections. The data were inconclusive about the most commonly used metadata standard for social media content. The most commonly reported metadata standards by both survey and interview participants was EAD and DACS which are the structure and content standards typically used in finding aids. This

aligns with the finding surrounding access where most repositories reported plans to use finding aids to support discoverability. Perhaps the metadata used to preserve social media content falls into the scope of study of digital preservation more generally because most archivists in this study agree that social media content does differ from other born-digital material: "Social media is not the end all be all of digital archives, and digital archiving. In all digital records you have issues of privacy, you have issues of custody, you have issues of authorship" (Project 2). Finding aids work well to support discoverability and support intellectual arrangement for collections that contain both analog and digital content. Most activist social media archiving projects in this study have not started widely publicizing their collections to receive feedback from researchers about their search experience.

The survey and interview results indicate that most archives have handled project management issues such as staffing and funding by absorbing their activist collections as part of their daily tasks. They usually feel that their initiative fits comfortably within their existing job descriptions and collecting missions meaning that they do not feel it puts a burden on their existing resources.

Of both survey and interview respondents, only Project 1 reported the likelihood of providing access to social media content in the reading room or via data hydration. This is largely because they collect using hashtags, where they have not received permission for every user they have content for. They explain:

> Because of Terms of Service we're very much likely to provide access to this stuff in a controlled manner… We could provide access in a controlled environment like the reading room—load data onto a laptop and make things available. There's this idea of hydration…You can provide a researcher with Tweet IDs and then they can go and hydrate those records with the full data that's available through the Twitter API. We couldn't make the Instagram content available on the open web except for in the cases

> where we got permission from the people who either took or posted the photos. (Project 1)

One aspect of access not pursued in the present study was inter-institutional discoverability of social media data. This could be an area of future exploration, as it was identified by Truman (2016) as a difficulty for researchers, especially when some data may only be available in a reading room.

5.3 Outcomes in donor and user engagement

Though many of the collecting initiatives described by respondents in the interviews and surveys are in the initial stages they reported some outcomes that reveal some possible impacts of activist social media archives. The most commonly reported outcome on the survey was media attention which helps to increase awareness of library activities. Two interview participants explain that outreach was a large part of the initial stages of the collecting initiative, which involved publicizing their projects and explaining the purpose.

Other valuable outcomes named during the interviews include an increase in the donation of materials and an increase in student interest in the archives. Both of these outcomes can be very valuable to an archive. Project 2 explained that though their activist collecting initiative focused on current student activism they received alumni donations from past campus activism. This indicates that these activist oriented collection initiatives may help document archival silences by encouraging the donation of related material from the past. Archives generally seek better student engagement with materials. If students are engaging more fully with these projects it may be because they are student focused, and about events in which they have already expressed interest.

5.4 Limitations and Areas of Future Research

Though this study revealed commonalities and differences in how repositories collect social media content, there are some limitations to its methodology. The limitations of this study are that its small sample size limits its generalizability. First, that the population is quite small. However, I received a good response rate considering that there are only a few activist social media collections. Secondly, collecting and analyzing qualitative data is time intensive, which limits the number of participants I could reasonably include. However, what qualitative data lacks in generalizability it makes up for in thoroughness from each participant. Another limitations is that coding qualitative data for analysis is an inherently subjective exercise. I offset any ambiguity introduced by the nature of the data by defining clear categories of analysis, and by coding each interview more than once to account for the natural changes that occurred as I became more familiar with the coding system I used.

Despite these limitations this study suggests areas of future research within the professional literature. These areas were described more fully in the above discussion. More concisely they are:

- Ways to identify and appraise social media content relevant to the collecting mission of a repository.
- Is informed consent for social media users required for large quantities of social media data?
- If not, how to ethically use and provide access to those datasets.
- Does a focus on activist collecting adequately fill archival silences or are they creating further documentation gaps.

- What are long-term effects of relying on outsourced services like Archive-It? Hopefully, a closer examination will make those services more robust, while also planning adequately for future eventualities.

Conclusion

Both data collection methods in this research study indicate that barriers to the collection of social media, and activist social media in particular, are not primarily technological. None of the interview participants reported challenges specific to harvesting or preserving social media data. Rather, it seems the challenges lie in the professional ethics surrounding providing long-term public access to content. It is more important, particularly for activist groups, to ensure that the data is used ethically so that it can be made accessible long into the future. Archivists working on activist collecting initiatives make informed decisions about the best way to ensure that social media content is only made accessible once the groups consent to its long-term preservation and use. In many ways, the archivists who participated in this study have ensured ethical research use by following a traditional provenance based approach to collecting social media, by only providing access to social media data that has been knowingly transferred to their repository. Yet the ambiguity remains for social media datasets that could be analyzed quantitatively. The uncertainty would be reduced if there was professional consensus about the ethical collection, preservation, and use of social media. As archivists on Project 1 explained, coming to this consensus likely will not only involve archivists but researchers as well. It is important to keep in mind that social media content differs only in format; it is not very different from other digital material that repositories regularly handle. As social media archives become more prevalent it will be

important to remember best practices in digital preservation generally, and not to reinvent the wheel when it comes to social media formats.

This study demonstrates that archivists use their professional knowledge to harvest, preserve, and provide access to social media content. However, the ethical collection development and use of activist social media still need to be addressed by the professional community.

References

Allen, E. (2013). "Update on the Twitter Archive at the library of Congress" *Library of Congress Blog.* http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/

Babbie, E. (2010). *The Practice of Social Science Research.* Wadsworth: Cengage Learning. Print.

Bastion, J., & Alexander, B. (2009). "Introduction: Communities and archives – symbiotic relationship." *Community Archives: Shaping of memory.* London: Facet Publishing, xxi-xxiv. Print.

Barrera-Gomez, J., & Erway, R. (2013). *Walk this way: Detailed steps for transferring born-digital content from media you can read in-house.* Dublin, Ohio: OCLC Research. http://www.oclc.org/content/dam/research/publications/library/2013/2013-02.pdf.

Bennett, W. L., & Segerberg, A. (2011). Digital Media and the Personalization of Collective Action. Information, Communication & Society, 14(6), 770–799. http://doi.org/10.1080/1369118X.2011.579141

Buck, et al. (2014). "Documenting Ferguson: Project explanation and Purpose." Retrieved from http://digital.wustl.edu/ferguson/DFP-Plan.pdf

Cadell, L. (2013). "Socially practical or practically unsociable? A study into social media policy experiences in Queensland cultural heritage institutions." *Australian*

Academic & Research Libraries, *44* (1), 3–13. http://doi.org/10.1080/00048623.2013.77385

Caswell, M. (2014). Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives.*Archival Science*, *14*(3-4), 307–322. http://doi.org/10.1007/s10502-014-9220-6

Collins, S., Durington, M., Daniels, G., Demyan, N., Rico, D., Beckles, J., & Heasley, C. (2013). Tagging Culture: Building a Public Anthropology through Social Media. *Human Organization*, *72*(4), 358–368. http://doi.org/10.17730/humo.72.4.v5x0205248427516

Cook, T. (2013). Evidence, memory, identity, and community: four shifting archival paradigms. *Archival Science*, *13*(2-3), 95–120. http://doi.org/10.1007/s10502-012-9180-7

Drake, J. (2015, December 2). Announcing ASAP: Archiving Student Activism at Princeton. Retrieved March 25, 2016, from https://blogs.princeton.edu/mudd/2015/12/announcing-asap-archiving-student-activism-at-princeton/

Erde, J. (2014). Constructing archives of the Occupy movement.*Archives and Records*, *35*(2), 77–92. http://doi.org/10.1080/23257962.2014.943168

Erway, R. (2010). "Defining born-digital." *OCLC Online Computer Library Center, Inc.* http://www.oclc.org/research/activities/hiddencollections/bornditgital.pdf

Farrell, J. (2016). "Archiving student action at HSL." Et Seq.: The Harvard Law School Library Blog. http://etseq.law.harvard.edu/2016/02/archiving-student-action-at-hls/

Flinn, A. (2007). Community Histories, Community Archives: Some Opportunities and Challenges. *Journal of the Society of Archivists,28* (2), 151–176. http://doi.org/10.1080/00379810701611936

Flinn, A. & Stevens, M. (2009). "'It is no mistri, wi mekin histri.' Telling our own story: Independent and community archives in the UK, challenging and subverting the mainstream." *Community Archives: Shaping of memory.* London: Facet Publishing, 3-27. Print.

Flinn, A., Stevens, M., & Shepherd, E. (2009). "Whose memories, whose archives? Independent community archives, autonomy and the mainstream." *Archival Science, 9* (1-2), 71–86. http://doi.org/10.1007/s10502-009-9105-2

Gilliland, Anne. (2008). "Setting the stage." *Introduction to Metadata*. Getty Publications.

Gruzd, Anatoliy, and Caroline Haythornthwaite. 2013. "Enabling Community Through Social Media."*Journal of Medical Internet Research* 15 (10). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3842435/.

Ham, F. G. (1975). "The Archival Edge." *The American Archivist, 38* (1), 4-13. http://americanarchivist.org/doi/pdf/10.17723/aarc.38.1.7400r86481128424

Harris, V. (2002). "The Archival Sliver: Power, Memory, and Archives in South Africa." *Archival Science, 2,* 63-86. Retrieved from http://www.nyu.edu/pages/classes/bkg/methods/harris.pdf

Haustein, S., T. D. Bowman, K. Holmberg, I. Peters, & V. Larivière. (2014). Astrophysicists on Twitter: An in-depth analysis of tweeting and scientific

publication behavior. *Aslib Journal of Information Management*, *66*(3), 279–296. http://doi.org/10.1108/AJIM-09-2013-0081

Higgins, S. (2008). The DCC Curation Lifecycle Model. *International Journal of Digital Curation*, *3*(1), 134–140. http://doi.org/10.2218/ijdc.v3i1.48

International Internet Preservation Consortium. (2012). "Web archiving." http://www.netpreserve.org/web-archiving/overview

Jeffrey, S. (2012). "A new Digital Dark Age? Collaborative web tools, social media and long-term preservation." *World Archaeology*, *44* (4), 553–570. http://doi.org/10.1080/00438243.2012.737579

Jules, B. (2015a). "Documenting the Now." *Medium.* Retrieved October 4, 2015, from https://medium.com/on-archivy/documenting-the-now-ferguson-in-the-archives-adcdbe1d5788

Jules, B. (2015b). "Hashtags of Ferguson." *Medium*. Retrieved fromhttps://medium.com/on-archivy/hashtags-of-ferguson-8f52a0aced87

Keough, B., & Schindler, A. C. (2004). "Thinking Globally, Acting Locally: Documenting Environmental Activism in New York State." *Archival Issues: Journal of the Midwest Archives Conference*, *28* (2), 121–135. Retrieved from https://auth.lib.unc.edu/ezproxy_auth.php?url=http://search.ebscohost.com/login.aspx?direct=true&db=lih&AN=22357864&site=ehost-live&scope=site

King, L. "Emory Digital Scholars Archive Occupy Wall Street Tweets." (2012). *Emory Report.* http://news.emory.edu/stories/2012/09/er_occupy_wall_street_tweets_archive/campus.html(October 7, 2015).

Kolachina, K. (2015). Student groups preserve their history with UCLA's new archive project. *Daily Bruin.* Retrieved March 25, 2016, from http://dailybruin.com/2015/04/10/student-groups-preserve-their-history-with-uclas-new-archive-project/

Krikorian, R. (2014). "Twitter #DataGrants selections" *Twitter Blog.* Retrieved October 8, 2015, from https://blog.twitter.com/2014/twitter-datagrants-selections

Liew, C., King, V., L., & Oliver, G. (2015). "Social Media in Archives and Libraries: A Snapshot of Planning, Evaluation, and Preservation Decisions." *Preservation, Digital Technology & Culture, 44* (1), 3–11. http://doi.org/10.1515/pdtc-2014-0023

Maryland Historical Society. (2015). "Announcing BaltimoreUprising2015.org." http://www.mdhs.org/announcing-baltimoreuprising2015org

Maycon, T. (2016). "Ithaca College president to resign following student, faculty backlash." *USA Today College.* http://college.usatoday.com/2016/01/14/ithaca-college-president-resigns/

McNealy, J. (2012). "The privacy implications of digital preservation: Social media archives and the social networks theory of privacy." *Elon Law Review, 3,* 133-160. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2027036.

Millar, L. (2010). *Archives: Principle and Practices.* London: Facet Publishing. Print.

Morning Edition. (2016). "Will Future Historians Consider These Days The Digital Dark Ages?" *NPR*. http://www.npr.org/2016/01/04/461878724/will-future-historians-consider-these-times-the-digital-dark-ages

NCSU Libraries (2015a). "Legal and Ethical Implications." *Social Media Archives Toolkit.* http://www.lib.ncsu.edu/social-media-archives-toolkit/legal.

----. (2015b). “Social Media Harvesting Tools.” *Social Media Archives Toolkit.* http://www.lib.ncsu.edu/social-media-archives-toolkit/collecting/social-media-harvesting-tools

NDSA. (2014). *The NDSA Content Working Group’s National Agenda Digital Content Area Discussions: Web and Social Media*. Retrieved from http://blogs.loc.gov/digitalpreservation/files/2014/01/NDSACWG_WebSocialMedia_Overview_Grotke.pdf

Pagowsky, N., & Wallace, N. (2015). “Black Lives Matter!: Shedding library neutrality rhetoric for social justice.” *C&RL News,* 196-200. Retrieved from http://crln.acrl.org/content/76/4/196.long

Paschild, C. N. (2012). Community Archives and the Limitations of Identity Considering Discursive Impact on Material Needs. *The American Archivist*, *75*(1), 125–142. Retrieved fromhttp://search.proquest.com.libproxy.lib.unc.edu/lisa/docview/1125212022/53DFA34DB8864A6CPQ/3?accountid=14244

Riedlmayer, A. & Naron, S. (2009). “From Yizkor Books to weblogs: genocide, grassroots documentation and new technologies.” *Community Archives: The Shaping of Memory.* London: Facet Publishing, 151-168. Print.

Roy Rosenzweig Center for History and New Media. (2011). “About.” *Occupy Archive.* www.occupyarchive.org/about (Retrieved on Oct. 8, 2015).

Sample, I. (2015). “Google boss warns of 'forgotten century' with email and photos at risk.” *The Guardian.* https://www.theguardian.com/technology/2015/feb/13/google-boss-warns-forgotten-century-email-photos-vint-cerf

Smith, D., & Thrasher, S. W. (2015). “Student activists nationwide challenge campus racism – and get results.” *The Guardian.* http://www.theguardian.com/us-news/2015/nov/13/student-activism-university-of-missouri-racism-universities-colleges

Society of American Archivists. (2013). *Describing Archives: A Content Standard, Second Edition.*

Springer, K. (2015). Radical Archives and the New Cycles of Contention.*Viewpoint Magazine.* Retrieved March 24, 2016, from https://viewpointmag.com/2015/10/31/radical-archives-and-the-new-cycles-of-contention/

Storrar, T. (2014). “Archiving Social Media.” *The National Archives blog.* https://web.archive.org/web/20150315215549/http://blog.nationalarchives.gov.uk/blog/archiving-social-media/

Squires, L. (2014). From TV Personality to Fans and Beyond: Indexical Bleaching and the Diffusion of a Media Innovation. *Journal of Linguistic Anthropology, 24*(1), 42–62. http://doi.org/10.1111/jola.12036

Summers, E. (2014). "On Forgetting and Hydration." *Medium*.. https://medium.com/on-archivy/on-forgetting-e01a2b95272**.**

Thomson, S. D. (2016). “Preserving social media.” *Digital Preservation Coalition.* http://dx.doi.org/10.7207/twr16-01

Truman, G. (2016). Web Archiving Environmental Scan. *Harvard Library Report.* Retrieved from https://dash.harvard.edu/handle/1/25658314

"Understanding Metadata." (2004). NISO Press. http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.

Wakimoto, D. K., Bruce, C., & Partridge, H. (2013). "Archivist as activist: lessons from three queer community archives in California." *Archival Science, 13* (4), 293–316. http://doi.org/10.1007/s10502-013-9201-1

Zach, L. & Peri, M. F. (2010). "Practices for college and university electronic records management (ERM) programs: Then and now." *The American Archivist, 73* (1), 105-128. http://www.jstor.org/stable/27802717.

Zhang, & Wildemuth. (2009). "Qualitative Analysis of Content." *Applications of Social Research Methods to Questions in Information and Library Science.* Connecticut: Libraries unlimited, 308-319. Print.

Zeng, D., Chen, H., Lusch, R. & Li, S. (2010). "Social Media Analytics and Intelligence." *IEEE Intelligent Systems, 25* (6), 13-16. Doi: 10.1109/MIS.2010.151.

Zinn, H. (1970). "The Archivist and Radical Reform." *Unpublished Manuscript.* http://www.libr.org/progarchs/documents/Zinn_Speech_MwA_1977.html

Appendix A: Archivists Survey

1. Were there already existing digital archival collections at your library?

   Yes
   No
   Don't know

   If so, what was it?

2. How was your social media collecting project funded?

   Grant
   It was free
   Money was allocated to it from the library or department budget
   Other

   If other please specify:

3. What social media platforms do you collect for your project?

   Facebook
   Twitter
   Tumblr
   Vine
   Vimeo
   Instagram
   YouTube
   Pinterest
   Flickr
   LinkedIn
   Blogs
   Other

   If other please specify:

4. What tools did you use to harvest social media?

Twarc
Lentil
Social Feed Manager
ArchiveSocial
ArchiveIT
Twitter Archiving Google Sheet (TAGS)
Personal Archives (Downloaded by individual users and donated to your collection)
Other

If other, please specify:

5. What were some challenges you faced regarding platform specific formats, or the tools used to collect data? Check any that apply:

   Legal issues (such as intellectual property, and privacy questions)
   Ethical issues (such as privacy, personal information, and intentions of the social media user)
   Harvesting data (tools were too difficult to use
   Available funds (tools and

6. What metadata do you use to help make your social media data more accessible? If you use a metadata standard what standard do you use? Check any that apply

   Dublin Core
   PREMIS
   Library Congress Subject Headings
   Folksonomies (such as user generated tags)
   MARC
   METS
   Other

   If other please specify:

7. What tools do you use to provide access to content? Check all that apply
   Omeka,
   ContentDM
   WordPress
   Homegrown tool
   Finding Aids
   Other

   If other please specify:

8. Name some of the outcomes you have experienced as a result of this project. Check all that apply.

   Wider use of analog and digital material
   Increase traffic to website or access platform
   Media attention
   Harassment
   Use in student projects
   Increased donations of material
   Increased monetary donations
   Increased attendance at library programs related to your social media collections
   Other

   If other please specify:

9. Is there any aspect of social media collecting that should be examined further by professionals and the professional literature?

10. Are you interested in being interviewed about your social media project?

    Yes
    No

Appendix B: Semi-Structured Interviews for Archivists

I. Work flows
   a. Why did you begin this project?
   b. How do collect social media content?
   c. How do you decide what to collect?
   d. What was important to you when making decisions about this project?
   e. How do you provide researchers with access? Can you describe the tools or process you use?
   f. Are there any issues you have had with providing patrons with access? If so have you created any policies to deal with these issues? What are they?
   g. On a fairly high level could you step me through the process of collecting social media for this process from start to finish?
   h. What were the greatest challenges you faced and how did you overcome them?
   i. What would have made these challenges easier for you?

II. Administrative concerns
   a. What were the campus wide policies that supported the creation of this project?
   b. Is there dedicated funding? How did you find funding?
   c. What other staff is involved with helping run the project if any?
   d. What is the future direction of this project or related projects?
   e. Describe one outcome of the project? Do you have a particular story you'd like to share about an outcome?

III. Cooperation and Coordination
   a. What stakeholders did you work with most closely on this project?
   b. What is your relationship with Library/Archives IT? How did they help you?
   c. What contact did you have with any legal team or institutional attorneys?
   d. Describe any legal issues you worked with them on if any
   e. What is your relationship with the activist group involved? How did they help you if at all?

IV. Best practices
   a. What best practices would you like to see implemented surrounding the collection and preservation of social media?
   b. Is there anything else you would like to share about collecting social media or working with activist groups?

V. If you're interested I would also like to talk to a member of the group that you have documented. Understanding the sensitive nature of the work that they do, I will keep their information as anonymous as possible.