Received: 15 August 2019 | Revised: 23 March 2020 | Accepted: 15 April 2020

DOI: 10.1002/asi.24368

RESEARCH ARTICLE

JASIST WILEY

Saving social media data: Understanding data management practices among social media researchers and their implications for archives

Libby Hemphill[1,2] | **Margaret L. Hedstrom**[1] | **Susan Hautaniemi Leonard**[2]

[1]School of Information, University of Michigan, Ann Arbor, Michigan, USA

[2]ICPSR, University of Michigan, Ann Arbor, Michigan, USA

Correspondence
Libby Hemphill, School of Information and ICPSR, University of Michigan, P.O. Box 1248, Ann Arbor, MI 48106-1248.
Email: libbyh@umich.edu

Funding information
National Science Foundation, Grant/Award Number: 1822228; IMLS, Grant/Award Number: RE-01-15-0086-15

Abstract

Social media data (SMD) offer researchers new opportunities to leverage those data for their work in broad areas such as public opinion, digital culture, labor trends, and public health. The success of efforts to save SMD for reuse by researchers will depend on aligning data management and archiving practices with evolving norms around the capture, use, sharing, and security of datasets. This paper presents an initial foray into understanding how established practices for managing and preserving data should adapt to demands from researchers who use and reuse SMD, and from people who are subjects in SMD. We examine the data management practices of researchers who use SMD through a survey, and we analyze published articles that used data from Twitter. We discuss how researchers describe their data management practices and how these practices may differ from the management of conventional data types. We explore conceptual, technical, and ethical challenges for data archives based on the similarities and differences between SMD and other types of research data, focusing on the social sciences. Finally, we suggest areas where archives may need to revise policies, practices, and services in order to create secure, persistent, and usable collections of SMD.

1 | INTRODUCTION

Social media are implicated in many of contemporary society's most pressing issues, from influencing public opinion, to organizing social movements, to identifying economic trends. Increasing the capacity of researchers to understand the dynamics of such phenomena will depend on reliable, curated, discoverable, and accessible social media data. To inform the development of research data infrastructure, we need to understand how researchers in this space work. This article reports on two related efforts to understand how researchers are using social media data for research and how they acquire and manage such data. We reviewed 40 papers in four journals that used data from Twitter to understand how authors described their research activities and then surveyed researchers about their social media data practices generally. Our goal was to inform the design of the social media archive (SOMAR) being developed at ICPSR, the oldest and one of the largest archives for managing and disseminating social science data, and to share our findings with other archives and research data management services as they incorporate this new data type.

We ask two different, but related, questions about the use of social media data for research: how do researchers use social media data in their research; and how do researchers acquire, manage, archive, and share social media data? We were not asking whether researchers are able to ask new questions, but rather analyzed research that uses Twitter data to gain an understanding of the types of questions researchers were using Twitter data to answer. We used this analysis to form a baseline idea of how people are using Twitter data in social science, and what kinds of questions they are asking.

We specifically address how social media researchers' practices may differ from what we know from previous studies of data practices, and we consider how the features of social media data (e.g., scale, speed, platform dependence, and ownership) influence data practices. We are particularly interested in the extent to which researchers' data management practices mirror (or don't) the data practices of researchers who use and share more traditional data types, such as surveys and administrative data. We discuss the properties of social media data, the types of research questions and methods reported in articles that rely on social media data, and the responses to our survey about data practices. We identify similarities and differences between social media data and more familiar types of data in order to discover gaps in current data archive models and identify where new approaches are needed because of some combination of the unique characteristics of social media data, the new approaches to social science research that they enable, and changing attitudes toward data management and data sharing.

1.1 | Twitter archiving

ICPSR is certainly not the first archive to consider processing and preserving social media data. Examples of other institutions, wrestling with the issues we raise, include GESIS, the Library of Congress, George Washington Libraries, the US National Archives, the UK Data Service, and the Documenting the Now Project, at a minimum. Researchers using data from social media platforms and practitioners developing archiving and dissemination services have raised issues about the scale and structure of data social media generate (Bruns & Weller, 2016; Zimmer, 2015), technical challenges in collecting data (Driscoll & Walker, 2014; Littman et al., 2018; Weller & Kinder-Kurlanda, 2015), ephemerality of social media data (Bruns & Weller, 2016; Littman et al., 2018; Weller & Kinder-Kurlanda, 2015; Zubiaga, 2018), platform application programming interface (API) restrictions (Bruns, 2019; Bruns & Weller, 2016; Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017; Littman et al., 2018; Thomson & Kilbride, 2015; Weller & Kinder-Kurlanda, 2015), challenges for documenting the provenance of social media data (Driscoll & Walker, 2014; Weller & Kinder-Kurlanda, 2015), and privacy concerns and the ethics around preserving and disseminating social media data (Fiesler & Proferes, 2018; Thomson & Kilbride, 2015; Wheeler, 2018; Zimmer, 2015). However, the relationship between these issues and researchers' data management practices needs further analysis to guide the development of effective approaches to preservation and reuse of social media data.

1.2 | Data sharing and management practices

Existing literature on researchers' data management practices tells us that although researchers are interested in sharing data, they rarely do so (Kennan & Markauskaite, 2015; Tenopir et al., 2011). Receiving credit for their work and maintaining the option or right to publish about the data first were important considerations for researchers when deciding whether to share their data (Wallis, Rolando, & Borgman, 2013). Many researchers negotiated private access to their data, especially within their groups, and with groups they knew and trusted, but were unwilling to share their data without restricting who could access the data and what scientific questions they were able to examine with it (Akers & Doty, 2013; Kennan & Markauskaite, 2015; Tenopir et al., 2011; Wallis, Rolando, & Borgman, 2013). They sometimes thought of data sharing as a "gift economy" in which they traded resources among trusted parties (Hilgartner & Brandt-Rauf, 1994; Wallis, Rolando, & Borgman, 2013), allowing them to barter for other resources in the process. Depositing data in an archive limits the bartering value of a particular data set, and the lack of credit, through data citation or other means, that researchers receive for sharing provides disincentive to do so.

Most researchers manage their data "privately" by storing it on local computers and hard drives (Akers & Doty, 2013; Whitmire, Boock, & Sutton, 2015). This local management practice was common even on campuses that offered secure, scalable storage, and computing resources through a centralized service (Whitmire, Boock, & Sutton, 2015). These practices mean that data are at risk for loss or leakage. Many datasets were not backed up in a second or secure storage space, placing them at risk for loss through both hardware failure and unauthorized access. Privately managed data are difficult for others to discover because they are hidden behind password-protected servers and file systems, not indexed or described to enable discovery, and subject to terms and conditions that are not available or transparent. According to the literature, researchers are also reluctant to share their data openly because they fear that the data will be misused or misinterpreted (Akers & Doty, 2013; Cragin, Palmer, Carlson, & Witt, 2010; Kim & Stanton, 2016). Effective data preservation depends, in part, on researchers' data management practices. Good data practices throughout the research lifecycle help ensure that users other than the original researchers will be able to find, understand, and reuse the data accurately (Goodman et al., 2014; Wilkinson et al., 2016). Requirements like data management plans, guidelines like Wilkinson's FAIR Principles (Findability, Accessibility, Interoperability, and Reusability), and standards for metadata and other types of documentation are intended to facilitate data management and data sharing, reduce the potential for misuse and misinterpretation, and ease the flow of data from researchers to permanent repositories. Nevertheless, research on data management practices and researchers' attitudes toward data sharing find that following the guidelines entails considerable effort and many researchers find adherence to such guidelines burdensome and time-consuming (Sayogo & Pardo, 2013; Tenopir et al., 2015). Earlier studies of sharing and management practices used surveys of broad populations of researchers (e.g., international [Tenopir et al., 2011], campus-wide [Akers & Doty, 2013; Whitmire, Boock, & Sutton, 2015]) or case studies of specific research centers and groups (e.g., [Mayernik, 2016; Wallis, Rolando, & Borgman, 2013]). Social science is particularly well represented in research on data management practices (Faniel, Kriesberg, & Yakel, 2016; Federer, Lu, Joubert, Welsh, & Brandys, 2015; Field et al., 2009; Kim & Adler, 2015; Pepe, Goodman, Muench, Crosas, & Erdmann, 2014). Social media, however, produce new types of data that researchers across a number of fields are using to address new questions. Little is known about research data management practices for social media data and few guidelines exist to assist researchers' selection and acquisition of data (Driscoll & Walker, 2014; Kinder-Kurlanda, Weller, Zenk-Möltgen, Pfeffer, & Morstatter, 2017). Our analyses of 40 peer reviewed publications presenting research that used data from Twitter and our survey of 73 researchers' data management practices were designed to gather insights into how social media data are used for research and what new data management challenges arise for researchers and for repositories like ICPSR that are developing guidance and services that will support this community most effectively.

2 | METHODS

We conducted two studies to better understand current practices among researchers who use social media data. First, we conducted a meta-analysis of articles that described acquiring, refining, and analyzing data from Twitter to gain insights into how researchers are using and analyzing social media data. Second, we surveyed social science researchers about their practices around collecting and sharing data from several social media applications to learn about their data management practices and needs.

We reviewed articles that appeared in four highly regarded interdisciplinary journals in order to effectively summarize current approaches to using social media data in research. In all, we reviewed 40 studies published in *First Monday*; *Information, Communication and Society*; *Journal of the Association of Information Science and Technology*; and *New Media & Society* (the full list of articles is available at https://www.openicpsr.org/openicpsr/project/109629/version/V2/view). We reviewed full papers, and coded each study according to its research question; data collection method (e.g., platform API and third-party provider); data set size; and sampling, statistics, and analysis approaches. We recognize that Twitter data are only one type of social media data and that research based on social media data are published in many other outlets, and we acknowledge that these four publications do not represent all of the disciplines that use social media data in research. We used these sources because the journals sit at the intersection of information science, computational science, and social sciences—core constituencies of ICPSR. We focus here on the data and analytic aspects of papers, details that were unavailable in earlier reviews of Twitter literature that reviewed only abstracts (Williams, 2013). We expected the breadth of disciplines and approaches reported in these journals to reveal a variety of methodological approaches to using Twitter data.

In our second study, we surveyed researchers who use social media data about their data management practices. We recruited respondents for our survey through email lists (e.g., AIR-L the listserv for the Association of Internet Researchers), Facebook groups (e.g., Researchers of the Socio-Technical), and investigators' individual social media accounts. We used this approach to better understand practices within a relatively narrow researcher population. Our sample does not allow us to generalize about social media data management practices but rather to uncover and articulate a range of practices and approaches. Our goal was to inform the development of archives and services around these data, so it was more important for us to understand the range of practices at this stage rather than their frequencies or representativeness relative to the entire population of researchers who use social media data. Data from the survey are available at https://www.openicpsr.org/openicpsr/project/109629/version/V2/view.

The survey was open from July 31, 2018, to August 21, 2018, and received 73 responses. Our survey instrument had five main sections: general and demographic, data acquisition, data transformation, analysis and visualization, and data sharing and reuse. We restricted our demographic data collection to an investigator's affiliation (e.g., university and

[Corrections added on 16 May 2020: An URL has been provided instead of text removed for blind review purpose in two instances.]

government lab) and position (e.g., PhD student, faculty, and staff) in order to focus on the researchers' practices rather than their individual characteristics. Prior work suggests that researchers in different age brackets and disciplines have different attitudes about data sharing (Wallis, Rolando, & Borgman, 2013), and we expect that some of those differences are also present in the population we surveyed.

Our current goal is to understand existing data management practices so that we and others who are building capacity to archive and disseminate social media data will be cognizant of current social media research practices, be able to identify common needs, and develop services that support researchers in data acquisition, management, archiving, and reuse. Together, these studies help us understand both research use and data management practices. We reserve more explicit questions about encouraging sharing of social media for future work.

3 | RESULTS

3.1 | Practices reported in publications

To understand the breadth of topics and research methods among social media researchers, we collected articles published in four interdisciplinary journals where researchers reported on empirical analyses of Twitter data. Overall, we did find variety in the topics covered, methods used, and scope and scale of studies in this sample of papers. We also found that most methods sections were (understandably) brief and did not provide rich detail about the data collection or transformation processes, and none of the studies provided access to their data or analysis in supplementary materials.

3.2 | Diversity of research areas

The authors of published articles used social media data to study a range of topics such as economic and consumer behavior (Antenucci, Cafarella, Levenstein, Ré, & Shapiro, 2014; Asur & Huberman, 2010), cultural differences (Hochman & Schwartz, 2012), social capital (Ellison, Vitak, Gray, & Lampe, 2014; Gil de Zúñiga, Jung, & Valenzuela, 2012), feminist and anti-racist movements (Brock, 2012; Dixon, 2014; Freelon, McIlwain, & Clark, 2016), political activism (Boulianne, 2015; Freelon, 2015; Roback & Hemphill, 2013), the relationship between social and traditional media (Jungherr, 2014; Papacharissi & de Fatima Oliveira, 2012; Shapiro & Hemphill, 2017; Soroka, Daku, Hiaeshutter-Rice, Guggenheim, & Pasek, 2018), and the impact and reach of research (Haustein et al., 2016; Thelwall, Haustein, Larivière, & Sugimoto, 2013). In our analysis of research that used Twitter data, we found a similar breadth of research topics, ranging from audience interactions around television shows (Boukes & Trilling, 2017; A. Williams & Gonlin, 2017) to social justice movements under hashtags such as #Ferguson (Barnard, 2018), and many political discussions around the world (Aelst, Erkel, D'heer, & Harder, 2017; Engesser, Ernst, Esser, & Büchel, 2017; Zelenkauskaite & Niezgoda, 2017; Zhang, Wells, Wang, & Rohe, 2017). Research topics were not limited to social and behavioral phenomena. Several studies used Twitter data to characterize social networks of users of particular hashtags (Rambukkana, 2015), to test Twitter's effectiveness as a communication medium (Coppock, Guess, & Ternovski, 2016; Gainous & Wagner, 2014), or to identify characteristics of tweets associated with concepts like trustworthiness or utility (Halse, Tapia, Squicciarini, & Caragea, 2018). The studies in our sample often relied on data acquired from third-party distributors rather than directly from Twitter. For instance, Crimson Hexagon and Radian6 were frequently mentioned. Data sets ranged in size from just over 100 images to over 2 million tweets. In some cases, the boundaries of the data set were established by content (e.g., hashtags and keywords) and in others by the authors of the content (e.g., members of parliament and journalists). Papers also reported a variety of analytical approaches requiring wide-ranging methodological and computational expertise (e.g., qualitative grounded theory and computationally intensive machine learning).

3.3 | Survey results

3.3.1 | Demographics and research areas

The vast majority of respondents (87.7%) are affiliated with universities, with faculty ($N = 23$) and PhD students

TABLE 1 Survey respondents' affiliations

Affiliation	% of respondents	*N*
University	87.7%	63
Faculty	31.5%	23
PhD student	23.3%	17
Master's student	12.3%	9
University Post-Doc	9.6%	7
Undergraduate student	5.5%	4
University staff	5.5%	4
Industry	6.9%	5
Government or non-profit	4.1%	3
Other or not indicated	1.4%	1
Total	100%	73

($N = 17$) making up more than half (54.8%) of all respondents (Table 1). Researchers in industry ($N = 5$) and government or non-profit organizations ($N = 3$) are not well represented in our survey, most likely because the types of email lists, online interest groups, and social networks we tapped for recruitment of subjects are more heavily populated with academic researchers.

We asked researchers whether the focus of their research was on some aspect of the use or users of social media platforms themselves (e.g., Facebook) or whether they analyzed user-generated content from social media platforms to understand some other phenomenon (e.g., economic trends). Thirty-eight of our 73 respondents (52%) chose "I study social media platforms and/or social media users themselves"; 17 (23%) chose "I use social media data to study something else beyond social media." Just six respondents chose "other" and supplied free-text answers that fell somewhere in between (e.g., "social media data as part of the agenda setting process") or said "both." Although the respondents, as a whole, used social media data from 11 different platforms (Table 2), very few reported collecting data from more than one platform.

3.3.2 | Data acquisition and analysis

We asked respondents to list tools or software they used to gather social media data. Python, the programming language, was the most frequent tool mentioned; and Python libraries such as pandas, scikit-learn, tensorflow, nltk, numpy, and related tools such as Jupyter notebooks were also mentioned. R or related tools (R Studio) were the next most frequent category of tools. Respondents who mentioned specific software or services listed NVivo, Discovertext, NodeXL, TAGS, IFTTT, Social Feed Manager, Zapier, Hydrator, WebRecorder.io, and SPSS. Eleven respondents (15%) said they had paid for access to social media data. We also asked respondents to indicate what skills they thought were important for people working with social media data to have. Their responses are summarized in Table 3.

Twenty-two respondents also provided an answer under "other" and indicated that skills such as "understanding of privacy issues/ethics of social media data," "thoughtful engagement with the ethics and accountability of their research," and "understanding of digital culture." Respondents also indicated that computational skills were not always necessary. For instance, one said, "I don't think any of these are 'necessary' as one can perform research on social media data via qualitative means," and another commented, "analytical skills, all the other things can come from a team."

When asked about where those skills were acquired, 63% of respondents ($N = 46$) said they had "learned on my own or with help online (e.g., Stack Overflow)". The options "taught by someone on my research team" and "platform API documentation" were both chosen by 27% of respondents ($N = 20$). Only 10% learned "in class" ($N = 7$). Other answers included "from a book" ($N = 11$), and "other" ($N = 7$). Among the "other" responses, people reported learning from colleagues, staff, and students who were not members of their research team.

3.3.3 | Data sharing and reuse

Twenty-three respondents (31.5%) said they do not make their data available to others. Thirty-four respondents (46.6%) who do make their data available use repositories and websites (Table 3). Eleven respondents chose "other" when asked "How do you make your data available to others?" In those responses, many mentioned restrictions on data sharing imposed by platforms or indicated that they would be willing to share data directly with researchers who asked. For instance, they indicated, "code is on GitHub, they can request data" or "they will receive an external hard drive with the data" and "We

TABLE 2 Social media platforms used to supply data for analysis

Platform	% of respondents	*N*
Twitter	39.7%	29
Facebook	28.8%	21
Instagram	11.0%	8
Reddit	11.0%	8
Wikipedia	6.8%	5
Tumblr	5.5%	4
Other	4.1%	3
Twitch	2.7%	2
YouTube	2.7%	2
Pinterest	1.4%	1

TABLE 3 Skills that respondents considered important

Skill	Respondents
Web scraping	38
Python	33
R	26
Advanced statistics	24
System/server administration	10

TABLE 4 Mechanisms used by respondents to share social media data

Mechanism	% of respondents	N
I make my data available.	46.6%	34
In a repository or archive	15.1%	11
Through a personal website	11.0%	8
Through journal or conference site	8.2%	6
Through a University-affiliated website	6.8%	5
Through a third-party data provider	5.4%	4
I don't make my data available.	31.5%	23
Other	15.1%	11
No response	6.8%	5

TABLE 5 Preparation and requests for reuse and replication

	Yes	No
Have you ever prepared your data especially for reuse?	21 (28.8%)	28 (38.4%)
Have you ever prepared your data especially for replication?	17 (23.3%)	38 (52.1%)
Has anyone ever contacted you, or your team, to request access to your social media data set?	11 (15.1%)	40 (54.8%)

can directly share signals we calculate from that data, but not the social media data itself" or "We make data available on a case-by-case basis, given platform Terms of Service." Respondents who used repositories or archives to share their data listed their university's institutional repositories ($N = 3$), Github ($N = 3$), Figshare ($N = 2$), and ICPSR ($N = 1$) (Table 4).

We also asked whether they had prepared data for reuse by anyone within their research groups ($N = 17$), by others outside their groups ($N = 14$), or not at all ($N = 28$). The majority of respondents had not received requests for their data or prepared their data for replication. Table 5 summarizes the results of these questions about preparation and requests for reuse or replication. When preparing for replication, respondents most often indicated that they provided code (e.g., Jupyter notebooks and R scripts) for analysis and filtered or cleaned datasets that contained only the data reported in a publication. When preparing data for sharing, respondents anonymized datasets, published tweet IDs, cleaned the data, and wrote documentation about their analysis process (e.g., README files).

3.4 | Summary of findings

Through our analysis of 40 papers that used Twitter data and our survey of researchers who use social media data, we reached three tentative conclusions. First, researchers used Twitter data to address a wide variety of issues ranging from characterizing the social networks of Twitter users to analyzing the content of tweets associated with particular hashtags, political issues, events, and other phenomena. The breadth of domains reflected in our data echoes that found by earlier reviews of Twitter research generally (Williams, 2013) and within the health sciences (Williams, Terras, & Warwick, 2013). Some of the studies we reviewed used Twitter data as a new source for insights into long-standing questions about social, behavioral, political, and economic issues, while other studies attempted to understand the impact of Twitter as a new form of communication. Second, our survey showed that using social media data for research requires more technical skills and familiarity with a wider variety of tools than research using more established sources, such as surveys, and methods, such as regression analysis. Most researchers gained these skills through informal means. It appears that a single individual rarely possesses the full complement of conceptual, analytical, computational, and technical skills needed to work with social media data; rather these skills are distributed across different members of research teams. Third, we found both similarities and differences between the data management and data sharing practices of researchers using social media data and what we know about other researchers from the literature. Researchers using social media data seem to focus their data management efforts on acquiring data and on making the data usable for their own analyses rather than on making the data reusable by others. We found that they raise concerns similar to those of other social scientists about sharing their data and ethical issues such as privacy and misinterpretation of data. Whether these differences are a consequence of unique characteristics of social media data, the new affordances of social media for novel paths of inquiry, the relative immaturity of social media research, or other factors is the topic of our discussion below.

4 | DISCUSSION

4.1 | What makes social media data different?

Social media data consist of user-generated content that users create, share, or react to, and system-generated data, such as timestamps, account information, and clickstreams. Typically, researchers acquire data directly from one or more social media platforms or submit requests to these private entities for data sets that meet specific criteria. The data are proprietary with differing terms of service depending on the platform of origin, which may place limits on researchers' requests to obtain access, customize data, link content to account information, share data with others, and archive the data. Social media data are updated constantly and usually delivered as raw feeds that generally require programming before analysis; historical data (sometimes as recent as two weeks old) are often more difficult or costly to access than live streams. Raw feeds consist of system-generated metadata (e.g., user account age and content creation date) and user-generated content (e.g., the text of a tweet or Facebook post), and pointers to resources that live elsewhere (e.g., photos, videos, and URLs). The platforms are unwilling to provide access to the proprietary algorithms that structure the streaming data into meaningful feeds.

4.2 | Data structures, scale, and speed

One challenge social media data present is the difficulty of describing what constitutes a "collection of social media data" or a "social media data set" (Voss, Lvov, & Thomson, 2017). Researchers and archives must know what it is they are proposing to collect, share, and archive, and the answer for social media data is not straightforward. One of our respondents commented that even our attempts to broadly define "data" and "data set" in the survey were too narrow: "You have a very limited concept of 'data set' underlying your questions. I collect live-streamed videos from protests." Should a social media data set include only the content from the social media platform (e.g., a tweet record from Twitter's API) or the social media content and the content it references (e.g., the contents of a URL included in a post, the video, or image shared)?

Platform terms of service also attempt to restrict what users of platform data can do with data they have collected, and researchers modify the data collected in order to comply with these terms (Bruns, 2019; Thomson & Kilbride, 2015). For instance, Twitter's Developer Policy—the agreement governing programmatic access to the site's content—states that people sharing Twitter content "will only distribute or allow download of Tweet IDs, Direct Message IDs, and/or User IDs" (Developer Policy, 2017). Does this then mean that Twitter datasets include only these items, and archives will be accepting and caring only for lists of identifiers rather than the content of the tweets? Tweets can be deleted from the platform at any time, by the author or by Twitter, and therefore, these shared lists of IDs are insufficient for reconstructing the original data sets. Research suggests that tweets in these ID collections persist at rates varying from 30% to 80% over four years (Zubiaga, 2018). Collections that contain only IDs are most likely incomplete because the objects they refer to are not persistent; for instance, tweets or posts may be deleted between when a researcher collects data and when they share that data with others. Respondents who used Twitter data and made efforts to share it for reuse or replication commonly reported sharing IDs—for example, "released tweet ids and jupyter notebooks"; "publishing tweet ids"; "all the tweet IDs". This means that even when researchers make efforts to share their data, the terms of service limit the completeness and replicability of their efforts.

Data from the articles we reviewed and the responses to our survey suggest researchers use different approaches to data collection (e.g., purchasing from third-party data resellers, and writing bespoke applications to collect data through APIs). Researchers then rarely describe the particulars of those collection methods or the transformations they perform on the data to prepare it for analysis. Instead, they report high-level efforts such as "cleaned data set, verified and cleaned code," or "I anonymized all the data upon collection" that do not detail the steps taken to clean data, or how researchers decided data was sufficiently anonymous. The inability to judge the quality or understand the provenance of a single research group's effort presents additional challenges for other research groups to reuse the data (Driscoll & Walker, 2014; Weller & Kinder-Kurlanda, 2016).

4.3 | Data practices: finding, curating, sharing, and storing data

Some fields have long histories of reusing structured survey, polling, observational, and administrative data, mature practices for managing such data, and experience with the reuse of data. For example, researchers understand that the design of a good survey includes documenting the sampling frames, data collection instruments, response rates, and measurement techniques by creating codebooks or using lab notebooks to keep track of the research process. (Wolf, Joye, Smith, & Fu, 2016). Although researchers have less control over the structure, quality, accuracy, and completeness of statistical, observational, and administrative data, they can use a combination of documentation, statistical techniques, and prior experience with canonical data sets (e.g., census data, economic indicators, and species registries) to detect errors or estimate reliability of data sets (Alvarez, 2016; Massey, Genadek, Alexander, Gardner, & O'Hara, 2018; Randall & Coast, 2016).

Sound data management practices, scalable curation, and archiving processes rely on documentation about the collection or creation of a data set, its internal structure, transformations performed on the data, use of field-specific ontologies, and metadata schema (e.g., The Open Biological and Biomedical Ontologies [OBO] Foundry, The National Library of Medicine's Repository of Common Data Elements [CDE], WordNet, SUMO), quality control measures (checking for completeness, validity of values, duplications, and adequate metadata), and the like (Goodman et al., 2014). When researchers create or collect their own data through surveys, interviews, experiments, and observation, they make choices about the quantity, structure, granularity, scope, and other aspects of the data as part of the research design. By documenting these decisions, data collections are more amenable to validation, replication, and reuse by others. Researchers also use administrative records, such as police reports, financial transactions, electronic health records, statistical compilations, and reference databases to address research questions. Unlike surveys, experiments, interviews, and observations, where researchers design a study and then create or collect data to address a particular research question, statistical, administrative, and other transactional data are not created explicitly for research. These types of data have been characterized as “found” (Harford, 2014; Mc Overton, Young, & Overton, 1993) or “non-designed” (Weinberg et al., 2019) data because they were not collected originally to address research questions. Rather, researchers discover data, assess its suitability for their research questions, and then manipulate the data for the specific purposes of their own research (Harford, 2014; Mc Overton, Young, & Overton, 1993).

Social media data are a new type of “found” data, and practices are evolving around new challenges that result from its use in research and its curation, dissemination, and reuse. Some of these practices are not unique to social media data: for example, reluctance to share data, difficulty in adding metadata, and risky data storage. Others, though, are more pronounced for social media data: for example, determining what constitutes a “collection” or “data set,” scaling methods of curation, and documenting data transformations. The processes of finding social media data and preparing it for use in research are frequently conducted computationally. Our respondents indicated that experience with computational skills such as programming, web scraping, and server administration are necessary for research that uses social media data. These skills are used at each stage of the data lifecycle—for example, Python scripts for collecting from the platform APIs, Jupyter, and R notebooks for cleaning and analyzing data. The computational processes involved in research with social media data present both challenges and opportunities for documenting workflow and preserving data provenance. Because the processes are captured in the code and/or notebooks, they are technically available for collection and preservation. However, most archives are not structured for or experienced with handling code and notebooks.

Researchers who use social media data showed a reluctance to share data for reasons that are similar to those expressed in other studies of researchers' attitudes toward data sharing (Tenopir et al., 2011; Whitmire, Boock, & Sutton, 2015). The resources, both computational and human, required to collect, transform, and manage social media data are non-trivial, and norms for recognizing this effort through citation, some share in authorship, or other means are nascent at best. Even when social media researchers are willing to share data upon request or distribute it through a website or repository, they are seeking guidance on how to document their data. No shared metadata standard for social media exists. Recent efforts by ad hoc groups of researchers have not gained traction (e.g., Open Collaboration Data Factories [OCDX-Specification, 2016]) nor produced proposals for metadata and documentation standards (e.g., Documenting Social Media Datasets [DocNow, n.d.], Datasheets for Datasets [Gebru et al., 2019]). These efforts and respondents' comments highlight that documenting social media data poses challenges in part because of the difficulty in describing the provenance of the data. For instance, the specific hashtags used to search for data through the Twitter API may change over the course of a project (e.g., a study of health care policy discussions begins by collecting #aca tweets, expands to include #obamacare, and #trumpcare tweets as those hashtags emerge). Documentation of the provenance of a social media data set should include the specific search terms, dates those terms were used, data returned that matched the query, and tracking of any subsequent transformations of the data, including the software and scripts used.

Finally, even among this computationally savvy group, researchers engage in risky data storage practices (e.g., using personal laptops instead of secured servers). Storing data on individual laptops increases risks of data loss and unauthorized access. Choosing to store locally rather than using secure data services is a common practice among academic researchers (Akers & Doty, 2013; Whitmire, Boock, & Sutton, 2015), and is not unique to social media data users. Though they eschewed university data services, many respondents reported using university license agreements for software (e.g., MaxQDA and NVivo).

4.4 | Ethical considerations in social media data management

Social media data also raise a host of new legal and ethical challenges. Private companies own and control the algorithms that underpin every aspect of how social media platforms operate, and they establish the terms and conditions for individuals who use these platforms in terms of personal privacy, proper use, intellectual property, and content limitations. Although platform users have some options for setting privacy and other use preferences, research has shown that privacy policies are ineffective at actually informing users about terms (Schaub, Balebako, & Cranor, 2017), and users make choices about sharing that depend on context (Acquisti, Brandimarte, & Loewenstein, 2015; Fiesler & Proferes, 2018). Social media users share sensitive and highly personal information, but it is unclear whether they are aware that this information could be harvested, archived, and reused without their explicit authorization. Our respondents recognized that social media data require special ethical consideration, including responses such as "understanding of privacy issues/ethics of social media data" and "thoughtful engagement with the ethics and accountability of their research" when asked about skills necessary for research with social media data. Responses to our survey also indicate that researchers who use social media data are seeking guidance on how to prevent disclosure of individual identities and sensitive information, protect privacy, and conform to unclear and sometimes contradictory ethical guidelines and contractual obligations. Existing research ethics guidelines generally focus on how to decide whether and how it is appropriate to use social media data in research (Fiesler & Proferes, 2018; Franzke, Bechmann, Zimmer, Ess, & the Association of Internet Researchers, 2020; Golder, Ahmed, Norman, & Booth, 2017; Townsend, 2017; Zimmer, 2010), and only recently have ethics guidelines for sharing and preserving that data emerged (Bishop & Gray, 2017; Weller & Kinder-Kurlanda, 2017). In addition to the familiar considerations of respect for persons, beneficence, and justice that often govern human subjects research, social media researchers, and archives must consider the legal implications of sharing and disseminating data that are "owned" in some sense by the for-profit platforms where it appears (Bishop & Gray, 2017; Bruns, 2019).

4.5 | Implications for archives

The breadth and diversity of practices present challenges for archiving, in part because the secondary uses may differ dramatically from the primary use of each data set. In addition, the context of reuse is fundamentally different from that of the social media platform where a user posted, responded to, or shared content originally. We discuss three ways in which social media differ data enough from the more familiar types of data that established archiving policies, and practices will need adjustment.

4.5.1 | Acquisition and manipulation of social media data

Most data archives acquire data sets either directly from a researcher or research team at the end of their project, or obtain data from administrative or statistical agencies on a regular cycle. Typically, these deposits include some documentation that explains how the data were acquired and organized into a data set or collection of data sets. Social media data, however, are first acquired by researchers from the social media platforms through their APIs or sites, or by way of special access negotiated with the platform providers, or through third-party distributors. All of these mechanisms for acquiring social media data place terms and conditions on what content and system-generated metadata can be downloaded, how the data can be used, and whether it can be shared with others. We learned from the survey that researchers use a variety of tools to acquire data and further manipulate the data to make it useful for their particular research questions. Placing restrictions on the conditions of use and reuse is not new to social media data, nor is the practice of cleaning and manipulating data prior to analysis. Nevertheless, it appears from our survey that researchers have greater challenges ascertaining the scope, depth, granularity, and temporality of the data they acquire from social media platforms and third parties, raising questions about the ability to benchmark social media data against some reality or ground truth. For instance, authors in our review who used services such as Crimson Hexagon and Netlytic did not describe how the third-party providers classify tweets, and different APIs return data that are incomplete or biased in different ways (Bruns, 2019; Driscoll & Walker, 2014; Morstatter, Pfeffer, Liu, & Carley, 2013). We also noted that the data are acquired and manipulated computationally. These new acquisition and research practices suggest that traditional notions of documentation may be inadequate, and that facilitating reuse of social media data by others will require much richer documentation of provenance, explicit documentation of the terms and conditions for acquiring the data, and documentation or deposit of the software and scripts used to acquire and manipulate the data.

4.5.2 | Technical and conceptual challenges

Social media data are complex objects that live in networks of relationships and linkages between user-generated content, metadata, external references, external content, and system-generated metadata. Compared to most types of archived data collections, social media data are especially voluminous and dynamic. For example, researchers may decide not to download linked content in order to comply with terms and conditions or for practical reasons, such as limiting storage requirements or improving the performance of the scripts used to scrape data from APIs. This means that linked content, which was available on the original platform, may have been deleted or changed by the time a researcher wishes to reuse the data. Current methods for curation are unlikely to scale for social media data, and they will remain ineffective and unaffordable without new tools and workflows for the currently laborious processes of metadata extraction and creation, quality control, and detection of disclosure risk (Voss, Lvov, & Thomson, 2017).

4.5.3 | Privacy, confidentiality, and ethical use of social media data

Established practices for informed consent, confidentiality and privacy protection, anonymization, and preventing deductive disclosure of individual identities are starting points for considering the ethical responsibilities that repositories incur when they acquire social media data. Nevertheless, new questions are arising about the appropriate use of social media data because of changing assumptions about consent, disclosure, persistence, and control over user-generated content. The terms and conditions for posting, sharing, and deleting content on social media platforms are governed by user agreements, platform terms of service, and individual configurations of privacy and other settings, as well as ever-changing norms about what is appropriate to post in the first place, who "owns" personal data, and how decisions are made about distribution, deletion, and disposition of social media data, and regulations such as the General Data Protection Regulation (GDPR) in the European Union (Mostert, Bredenoord, Biesaart, & van Delden, 2016; Politou, Alepis, & Patsakis, 2018).

The results of our survey suggest that researchers are seeking guidance on many of the issues we have discussed. Tackling this complex challenge while building on the knowledge and experience of both researchers and curators will require collaboration between repositories, such as SOMAR and GESIS, that are developing new archiving capacity for social media data and researchers, who are encountering myriad conceptual, technical, and ethical questions as they bring innovative methods and new types of data sources into their research. It is worth noting that in our survey students constitute the largest single group engaged in research using social media data; they are also the most frequent users of ICPSR data. Aiming services and training at students at the beginning of their careers may be more effective than trying to reeducate more senior scholars with entrenched habits.

5 | CONCLUSION

Research that relies on data from social media covers a wide range of topics, allows new research questions to be formulated and addressed, and creates opportunities to address old questions in novel ways. The data management practices employed for working with social media data resemble the processes for many other types of data, especially other types of "found" data such as censuses, statistical compilations, administrative records, and records of financial transactions. However, for other found data, documentation and storage standards are generally agreed upon, and data archives around the globe offer guidance for researchers working with such data. Standards for social media data are nascent, and archives are just beginning to offer support.

Researchers who use social media data also mirror other researchers in their reluctance to share data without ensuring credit for their work, awareness of who will reuse the data, and confidence that the data will not be used inappropriately. Social media data are an uneasy fit in existing data archives due to differences in scale, speed, platform dependence, structure, and ownership. An archive that facilitates the preservation and reuse of social media data will need to contend with additional challenges in documenting data and its provenance, in describing what constitutes a "dataset" in this space, and in ensuring appropriate protections for personal and sensitive information.

ACKNOWLEDGMENTS

We are grateful to student researchers Rebekah Small, Joshua Guber man, and Saul Hank in for their assistance. This material is based upon work supported by the National Science Foundation under Grant No.1822228. This research was made possible in part by a grant from the United States Institute of Museum and Library Services, Laura Bush 21st Century Librarian Program, 'Research Experience for Masters Students', #RE-01-15-0086-15.

ORCID

Libby Hemphill https://orcid.org/0000-0002-3793-7281
Margaret L. Hedstrom https://orcid.org/0000-0002-0356-6806
Susan Hautaniemi Leonard https://orcid.org/0000-0003-2732-1341

REFERENCES

Acquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science*, *347*(6221), 509–514.

Aelst, P. V., Erkel, P. v., D'heer, E., & Harder, R. A. (2017). Who is leading the campaign charts? comparing individual popularity on old and new media. *Information, Communication & Society*, *20*(5), 715–732.

Akers, K. G., & Doty, J. (2013). Disciplinary differences in faculty research data management practices and perspectives. *International Journal of Digital Curation*, *8*(2), 5–26.

Alvarez, M. R. (2016). Introduction. In Alvarez, M. R. (Ed.), *Computational social science: Discovery and prediction* (pp.1–25). New York: Cambridge University Press.

Antenucci, D., Cafarella, M., Levenstein, M., Ré, C., & Shapiro, M. D. (2014). *Using social media to measure labor market flows* (No. 20010).

Asur, S., & Huberman, B. A. (2010). *Predicting the future with social media*. In 2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology, Vol. 1, pp. 492–499.

Barnard, S. R. (2018). Tweeting #Ferguson: Mediatized fields and the new activist journalist. *New Media & Society*, *20*(7), 2252–2271. https://doi.org/10.1177/1461444817712723

Bishop, L. and Gray, D. (2017), Ethical Challenges of Publishing and Sharing Social Media Research Data, Woodfield, K. (Ed.), The Ethics of Online Research (Advances in Research Ethics and Integrity, Vol. 2, pp. 159-187). UK: Emerald Publishing Limited, https://doi.org/10.1108/S2398-601820180000002007

Boukes, M., & Trilling, D. (2017). Political relevance in the eye of the beholder: Determining the substantiveness of TV shows and political debates with twitter data. *First Monday*, *22*(4).

Boulianne, S. (2015). Social media use and participation: a meta-analysis of current research. *Information, Communication & Society*, *18*(5), 524–538.

Brock, A. (2012). From the blackhand side: Twitter as a cultural conversation. *Journal of Broadcasting & Electronic Media*, *56*(4), 529–549.

Bruns, A. (2019). After the "APIcalypse": social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Bruns, A., & Weller, K. (2016). *Twitter as a first draft of the present: and the challenges of preserving it for the future*. In Proceedings of the 8th ACM conference on web science, pp. 183–189. ACM.

Coppock, A., Guess, A., & Ternovski, J. (2016). When treatments are tweets: A network mobilization experiment over twitter. *Political Behavior*, *38*(1), 105–128.

Cragin, M. H., Palmer, C. L., Carlson, J. R., & Witt, M. (2010). Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *368*(1926), 4023–4038.

Developer Policy. (2017). Retrieved from https://developer.twitter.com/en/developer-terms/policy.html.

Dixon, K. (2014). Feminist online identity: Analyzing the presence of hashtag feminism. *Journal of Arts and Humanities*, *3*(7), 34–40.

DocNow. (n.d.). Retrieved from https://www.docnow.io/.

Driscoll, K., & Walker, S. (2014). Big data, big questions—working within a black box: Transparency in the collection and production of big twitter data. *International Journal of Communication Systems*, *8*, 20.

Ellison, N. B., Vitak, J., Gray, R., & Lampe, C. (2014). Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. *Journal of Computer-Mediated Communication*, *19*(4), 855–870.

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2017). Populism and social media: how politicians spread a fragmented ideology. *Information, Communication & Society*, *20*(8), 1109–1126.

Faniel, I. M., Kriesberg, A., & Yakel, E. (2016). Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology*, *67*(6), 1404–1416.

Federer, L. M., Lu, Y.-L., Joubert, D. J., Welsh, J., & Brandys, B. (2015). Biomedical data sharing and reuse: Attitudes and practices of clinical and scientific research staff. *PLoS One*, *10*(6), e0129506.

Field, D., Sansone, S.-A., Collis, A., Booth, T., Dukes, P., Gregurick, S. K., … Wilbanks, J. (2009). omics data sharing. *Science*, *326*(5950), 234–236.

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of twitter research ethics. *Social Media + Society*, *4*(1), 1–14. https://doi.org/10.1177/2056305118763366

Franzke, A. S., Bechmann, A., Zimmer, M., Ess, C., & the Association of Internet Researchers. (2020). *Internet research: Ethical guidelines 3.0* (Technical Report).

Freelon, D. (2015). Discourse architecture, ideology, and democratic norms in online political discussion. *New Media Society*, *17*(5), 772–791.

Freelon, D., McIlwain, C. D., & Clark, M. (2016). *Beyond the hashtags: #ferguson, #blacklivesmatter, and the online struggle for offline justice*. Washington, D.C.: Center for Media and Social Impact, American University. http://dx.doi.org/10.2139/ssrn.2747066

Gainous, J., & Wagner, K. M. (2014). *Tweeting to power: The social media revolution in american politics*. New York: Oxford University Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2019). *Datasheets for datasets*. http://arxiv.org/abs/1803.09010

Gil de Zúñiga, H., Jung, N., & Valenzuela, S. (2012). Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of Computer-Mediated Communication*, *17*(3), 319–336.

Golder, S., Ahmed, S., Norman, G., & Booth, A. (2017). Attitudes toward the ethics of research using social media: A systematic review. *Journal of Medical Internet Research*, *19*(6), e195.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data. *PLoS Computational Biology*, *10*(4), e1003542.

Halse, S. E., Tapia, A., Squicciarini, A., & Caragea, C. (2018). An emotional step toward automated trust detection in crisis social media. *Information, Communication & Society*, *21*(2), 288–305.

Harford, T. (2014). Big data: A big mistake? *Significance*, *11*(5), 14–19.

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & Larivière, V. (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on twitter. *Journal of the Association for Information Science and Technology*, *67*(1), 232–238.

Hilgartner, S., & Brandt-Rauf, S. I. (1994). Data access, ownership, and control: Toward empirical studies of access practices. *Knowledge*, *15*(4), 355–372.

Hochman, N., & Schwartz, R. (2012). Visualizing instagram: Tracing cultural visual rhythms. In *Proceedings of the workshop on social media visualization (SocMedVis) in conjunction with the sixth international AAAI conference on weblogs and social media (ICWSM–12)*, pp. 6–9.

Jungherr, A. (2014). The logic of political coverage on twitter: Temporal dynamics and content. *The Journal of Communication*, *64* (2), 239–259.

Kennan, M. A., & Markauskaite, L. (2015). Research data management practices: A snapshot in time. *International Journal of Digital Curation*, *10*(2), 69–95.

Kim, Y., & Adler, M. (2015). Social 'scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management*, *35*(4), 408–418.

Kim, Y., & Stanton, J. M. (2016). Institutional and individual factors affecting scientists' data-sharing behaviors: A multilevel analysis. *Journal of the Association for Information Science and Technology*, *67*(4), 776–799.

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), 2053951717736336.

Littman, J., Chudnov, D., Kerchner, D., Peterson, C., Tan, Y., Trent, R., … Wrubel, L. (2018). API-based social media collecting as a form of web archiving. *International Journal on Digital Libraries*, *19*(1), 21–38.

Massey, C. G., Genadek, K. R., Alexander, J. T., Gardner, T. K., & O'Hara, A. (2018). Linking the 1940 U.S. census with modern data. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, *51*(4), 246–257.

Mayernik, M. S. (2016). Research data and metadata curation as institutional issues. *Journal of the Association for Information Science and Technology*, *67*(4), 973–993.

Mc Overton, J. C., Young, T. C., & Overton, W. S. (1993). Using 'found' data to augment a probability sample: Procedure and case study. *Environmental Monitoring and Assessment*, *26*(1), 65–83.

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the sample good enough? comparing data from twitter's streaming API with twitter's firehose. In *ICWSM*. Retrieved from aaai.org.

Mostert, M., Bredenoord, A. L., Biesaart, M. C. I. H., & van Delden, J. J. M. (2016). Big data in medical research and EU data protection law: Challenges to the consent or anonymise approach. *European Journal of Human Genetics*, *24*(7), 956–960.

OCDX-Specification. (2016). Retrieved from https://github.com/OCDX/OCDX-Specification.

Papacharissi, Z., & de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on# egypt. *The Journal of Communication*, *62*(2), 266–282.

Pepe, A., Goodman, A., Muench, A., Crosas, M., & Erdmann, C. (2014). How do astronomers share data? reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLoS One*, *9*(8), e104798.

Politou, E., Alepis, E., & Patsakis, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cyber Security*, *4*(1), 1–20. http://dx.doi.org/10.1093/cybsec/tyy001

Rambukkana, N. (2015). *Hashtag publics: The power and politics of discursive networks*. New York: Peter Lang.

Randall, S., & Coast, E. (2016). The quality of demographic data on older africans. *DemRes*, *34*, 143–174.

Roback, A., & Hemphill, L. (2013). I'd have to vote against you: issue campaigning via twitter. In Proceedings of the 2013 conference on computer supported cooperative work companion (pp. 259–262). New York, NY: ACM.

Sayogo, D. S., & Pardo, T. A. (2013). Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly*, *30*, S19–S31.

Schaub, F., Balebako, R., & Cranor, L. F. (2017). Designing effective privacy notices and controls. *IEEE Internet Computing*, 1–1. http://dx.doi.org/10.1109/MIC.2017.265102930

Shapiro, M. A., & Hemphill, L. (2017). Politicians and the policy agenda: Does use of twitter by the U.S. congress direct new york times content? *Policy & Internet*, *9*(1), 109–132.

Soroka, S., Daku, M., Hiaeshutter-Rice, D., Guggenheim, L., & Pasek, J. (2018). Negativity and positivity biases in economic news coverage: Traditional versus social media. *Communication Research*, *45*(7), 1078–1098.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLoS One*, *6*(6), e21101.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., … Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS One*, *10*(8), e0134826.

Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? twitter and ten other social web services. *PLoS One*, *8*(5), e64841.

Thomson, S. D., & Kilbride, W. (2015). Preserving social media: The problem of access. *New Review of Information Networking*, *20* (1–2), 261–275.

Townsend, L. (2017). The ethics of using social media data in research: A new framework. In C. Wallace & W. Kandy (Eds.), *The ethics of online research* (Vol. 2, pp. 189–207). Bingley, UK: Emerald Publishing Limited.

Voss, A., Lvov, I., & Thomson, S. D. (2017). Data storage, curation and preservation. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 161–176). London: SAGE Publications Ltd.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology. *PLoS One*, *8*(7), e67332.

Weinberg, D. H., Abowd, J.M., Belli, R. F., Cressie, N., Folch, D. C., Holan, S. H., … Wikle, C. K. (2019). Effects of a Government-Academic Partnership: Has the NSF-CENSUS BureauResearch Network Helped Improve the US Statistical System? *Journal of Survey Statistics and Methodology*, *7*(4), 589–619. https://doi.org/10.1093/jssam/smy023

Weller, K., & Kinder-Kurlanda, K. (2017). To share or not to share? ethical challenges in sharing social media-based research data. In M. Zimmer & K. Kinder-Kurlanda (Eds.), *Internet research ethics for the social age: New challenges, cases, and contexts* (pp. 115–129). New York: Peter Lang Publishing, Incorporated.

Weller, K., & Kinder-Kurlanda, K. E. (2015). Uncovering the challenges in collection, sharing and documentation: The hidden data of social media research. In *Standards and practices in large-scale social media research. oxford: International conference on web and social media*.

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. In *Proceedings of the 8th ACM conference on web science* (pp. 166–172). ACM.

Wheeler, J. (2018). Mining the first 100 days: Human and data ethics in twitter research. *Journal of Librarianship and Scholarly Communication*, *6*(2), eP2235. http://dx.doi.org/10.7710/2162-3309.2235

Whitmire, A. L., Boock, M., & Sutton, S. C. (2015). Variability in academic research data management practices: Implications for data services development from a faculty survey. *Programmirovanie*, *49*(4), 382–407.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Williams, A., & Gonlin, V. (2017). I got all my sisters with me (on black twitter): second screening of how to get away with murder as a discourse on black womanhood. *Information, Communication & Society*, *20*(7), 984–1004.

Williams, S. A. (2013). What do people study when they study twitter? classifying twitter related academic papers. *Journal of Documentation*, *69*(3), 384–410.

Williams, S. A., Terras, M., & Warwick, C. (2013). How twitter is studied in the medical professions: A classification of twitter papers indexed in PubMed. *Med 2 0*, *2*(2), e2.

Wolf, C., Joye, D., Smith, T. W., & Fu, Y.-C. (2016). *The SAGE handbook of survey methodology*. London: SAGE Publications.

Zelenkauskaite, A., & Niezgoda, B. (2017). "Stop kremlin trolls:" Ideological trolling as calling out, rebuttal, and reactions on online news portal commenting. *First Monday*, *22*(5).

Zhang, Y., Wells, C., Wang, S., & Rohe, K. (2017). Attention and amplification in the hybrid media system: The composition and activity of donald trump's twitter following during the 2016 presidential election. *New Media & Society*, *20*(9), 3161–3182.

Zimmer, M. (2010). "But the data is already public": On the ethics of research in facebook. *Ethics and Information Technology*, *12*(4), 313–325.

Zimmer, M. (2015). The twitter archive at the library of congress: Challenges for information practice and information policy. *First Monday*, *20*(7).

Zubiaga, A. (2018). A longitudinal assessment of the persistence of twitter datasets. *Journal of the Association for Information Science and Technology*, *69*(8), 974–984.

How to cite this article: Hemphill L, Hedstrom ML, Leonard SH. Saving social media data: Understanding data management practices among social media researchers and their implications for archives. *J Assoc Inf Sci Technol*. 2021;72:97–109. https://doi.org/10.1002/asi.24368