ISSN 2162-3309 10.7710/2162-3309.2235

PRACTICE

Mining the First 100 Days: Human and Data Ethics in Twitter Research

Jonathan Wheeler
Data Curation Librarian, University of New Mexico

INTRODUCTION This case study describes data collection from Twitter, Inc. conducted with the intent of capturing conversations following from President Trump's and others' use of the #MAGA ("Make America Great Again") hashtag in Twitter posts during the first 100 days of his presidential administration. **DESCRIPTION OF PROGRAM** Data was collected between November 2016 and May 2017, using Twitter's public search, user timeline, and streaming APIs. **NEXT STEPS** The article discusses the ethical implications of collecting data from Twitter and describes the impact of Twitter's terms of service and API policies on data collection and research. **IMPLICATIONS FOR PRACTICE** Librarians engaged with data literacy and research conduct programs can support researchers in developing awareness of the context sensitivities of social media research. Data librarians and others involved with data management planning can where applicable provide guidance and resources to support ethical social media data collection and management. Twitter and other social media datasets which may be published within library supported institutional or data repositories must meet policy requirements.

Received: 01/15/2018 Accepted: 03/26/2018

Correspondence: Jonathan Wheeler, The University of New Mexico, Centennial Library L172, MSC05 3020, 1 University of New Mexico, Albuquerque, NM 87131, jwheel01@unm.edu

INTRODUCTION

Following the 2016 U.S. presidential election and prior to the January 20, 2017, inauguration of Donald Trump, the author and a colleague initiated a study of Twitter to investigate how different audiences receive and react to populist messages and trends. Specifically, our intent was to analyze what Twitter users "heard" and how they responded to Trump's use of the #MAGA ("Make America Great Again") hashtag through the first 100 days of the new administration.

The selection of Twitter as the platform of study and data collection was driven by two factors. From a practical standpoint, Twitter is Trump's platform of choice for communicating with his base. The conversations, trends, and follower networks of interest are therefore largely originating and taking place within a Twitter context. Further, the default public status of large amounts of Twitter data appealed to us as librarians and researchers. Issues related to big data collection, analysis, and curation are increasingly relevant to our work in information literacy and research data management. The opportunity to engage with these issues in practice has proven beneficial to our outreach, instruction, and consultation in these areas.

However, this was for both of us the first Twitter or social media research (SMR) project of any scale. As nonexperts we anticipated various technical challenges, but as the research has proceeded, standard questions around data collection and analysis methodologies—questions about the *process* of SMR—have been supplemented by deeper philosophical questions about the *nature* and *conduct* of SMR. As discussed in the lessons learned section below, as moderate social media users we were sensitive to the ethical contexts of Twitter research. However, it was during data collection and summary analysis that the scope of these issues at scale became apparent. Aside from strictly practical concerns about data management, more immediate and problematic issues of human agency, data ownership, and data ethics adhere to SMR regardless of the scope or scale of any particular dataset.

For example, relevant concerns include the nature of Twitter data as a corporate asset and the proliferation of social media studies in which the apparent simplicity of data collection clouds concerns about Twitter users as human subjects potentially governed by Institutional Review Board (IRB) research requirements. By way of elaborating on these issues, without going into content analysis we here present our data collection as a case study which demonstrates the impact of ethical and technological constraints on the access and use of Twitter data. These constraints delineate conflicts between personal, academic, and corporate interests that impact the agency of social media users, while also illustrating the effect of corporate control over data that has significant personal and research value. Implications for

scholarly communications librarianship as discussed below include addressing the ethical ramifications of SMR in data literacy programming and data management planning, as well as identifying limitations on data sharing and reproducibility resulting from Twitter policy.

LITERATURE REVIEW

There is a large and growing body of literature on social media and Twitter research. For a big picture overview of the SMR ecosystem and its evolution, recommended texts include *The Culture of Connectivity: A Critical History of Social Media* (Van Dijck, 2013), and *Twitter and Society* (Weller, Bruns, Burgess, Mahrt, & Puschmann, 2014). An extensive online bibliography is maintained by danah boyd (Bibliography of Research on Twitter & Microblogging, n.d.). Specific to the issues addressed in this paper, the literature reviewed here is focused first on evolving identities of self and community among social media users and the ethical implications for research relating to privacy and consent. A second section covers literature treating the data collection concerns associated with Twitter research and the platform's de-facto privileging of certain uses and users via its terms of service (TOS) and application programming interface (API).

Ethics: Changing Identities, Changing Contexts

Twitter exemplifies the ways in which social media has blurred the boundaries between information producers and consumers. This fluid dynamic is evident not only in the various contexts in which Twitter is leveraged by users to define new personal and public spaces, but also in the increasingly broad application of Twitter as a data source for research across disciplinary contexts.

Several case studies illustrate the ways in which Twitter has given rise to a citizen journalism that challenges the agenda and authority of mainstream media. For example, in their analysis of the social media response to the extrajudicial 2014 execution of Michael Brown, Bonilla and Rosa (2015) describe how strategic Twitter use by Ferguson residents and African Americans nationwide elevated the incident into the media spotlight through targeted use of the #Ferguson hashtag. Notably, Twitter use in this and other cases went beyond bringing attention to a single event and provided a forum for challenging racial and media stereotypes in broader contexts (Bonilla & Rosa, 2015, p. 5). Use of Twitter to call attention to issues and follow-on conversations in this fashion is not uncommon. For example, although Kalsnes, Krumsvik, and Storsul (2014) don't report a direct challenge to the dominant media agenda by Twitter users during Norwegian political debates, they do note a trend of Twitter commentary or "meta talk" ranging from assessments of candidate responses to critiques of their appearance (Kalsnes et al., 2014, p. 319). The implications of

Twitter's emergence as a real-time news source giving voice to locals and participants in major events is similarly addressed by Risse, Peters, Senellart, and Maynard (2014), and Crawford and Finn (2015).

Schmidt (2014) in particular illustrates the impact of this broadening of journalistic authority in his elaboration of the rise of "personal publics" on Twitter. In contrast to prior norms of newsworthiness and "publicness" as traditionally understood, Schmidt defines the personal public as a media space wherein information is selected and shared for a particular, personal audience (Schmidt, 2014, p. 4) and news has become participatory and conversational (Schmidt, 2014, p. 10). As implied by the case studies referenced above, this redefined personal public extends to a "communal public" relevant to ethical considerations of privacy and consent in Twitter research where there is no longer a clear distinction between public and private.

Users' reasonable expectations regarding data access and use is therefore a critical context in Twitter research. Zimmer and Proferes (2014a), Taylor and Pagliari (2017), and Robbins (2017) discuss from multiple perspectives the nuances of public information in social media and the ethical risks of assuming that "default public" platforms like Twitter necessarily carry a minimal expectation of user privacy. In addition to privacy as a factor of a user's intended audience, perceptions of privacy on Twitter are also informed by the supposed ephemerality of tweets. Not only do Twitter users have a constructed or curated public, but as discussed in depth by Zimmer and Proferes (2014a), Twitter executives' own rhetoric about the platform and features of the platform itself reinforce user expectations that tweets are short lived. This perceived ephemerality begs the question of whether users would communicate differently given the potential for tweets to be stored, preserved, or accessible over time.[1]

Privacy on Twitter, then, is not an absolute matter of excluding information from public view, but rather a context dependent assessment founded on users' expectations about whom they are sharing their tweets with and for how long. Users who perceive a narrow audience for their status updates may share information which they would not otherwise make public, a consideration that is particularly relevant to vulnerable populations. Golder, Ahmed, Norman, and Booth (2017) identify children, adults with medical conditions, and homosexuals as groups whose online activities may carry a greater risk of harm. Not only may standard considerations of privacy be insufficient to protect Twitter users within these populations, but Crawford and Finn (2015) provide further discussion

[1] The current versions of Twitter's terms of service and Privacy Policy do include specific language regarding data storage and third-party access.

of risk and the potential for discrimination against vulnerable populations when different social media datasets are combined.

Concern for vulnerable populations within Twitter research is further highlighted by the possibility that a user whose activity is captured within any given dataset has publicly retweeted the private tweet or message of another user (Mao, Shuai, & Kapadia, 2011; Zimmer & Proferes, 2014a). A 2011 case study by Mao et al. demonstrates multiple ways in which certain types of tweets or retweets can violate the privacy of Twitter users and others. In their study of privacy leaks on Twitter, the authors identified three classes of tweets—vacation tweets, drunk tweets, and disease tweets—which can respectively put users at risk of surveillance by burglars, law enforcement, employers, or insurance companies (Mao et al., 2011, p. 1). The authors further document not just particular types of leaks but also the sources of leaks—that is, whether specific classes of private information are more likely to be tweeted by users themselves through a status update or by their friends in a user mention (Mao et al., 2011). Some privacy leaks such as vacation information may pose only a temporary risk, but patterns of alcohol and drug use as well as disease or medical information may be evident in the social media activity of vulnerable populations and can remain sensitive over time. Such information may well be contained within a dataset.

Consent is another aspect of SMR in which assumptions based on past practice are subject to evaluations of context and user intent. Specific to Twitter, on the surface there is some argument to be made that acceptance of Twitter's Terms of Service (Twitter, Inc., 2017c) both limits a user's expectation of privacy and constitutes consent for ad hoc use of their public content. The broadcast and default public status of tweets carries implications as noted by the Privacy Policy (Twitter, Inc., 2017b). Specifically, the Privacy Policy explicitly states that "What you share on Twitter may be viewed all around the world instantly" (Twitter, Inc., 2017b). Additional language in the Privacy Policy refers to academic and market research as data use cases. Tweets and some user profile information are among content types defined as public information intended for distribution:

> Twitter broadly and instantly disseminates your public information to a wide range of users, customers, and services, including search engines, developers, and publishers that integrate Twitter content into their services, and organizations such as universities, public health agencies, and market research firms that analyze the information for trends and insights (Twitter, Inc., 2017b, Information Collection and Use section, para. 4).

Similarly, while the Twitter Developer Agreement and Policy (Twitter, Inc., 2017a) re-

quires user consent for certain activities, the relevance of the specified activities to research is unclear. For example, user consent is required to "Republish Content accessed by means other than via the Twitter API or other Twitter tools" (Twitter, Inc., 2017a, Developer Policy section 1.C.1.b). The API and tools developed using API data constitute an ecosystem of applications designed to republish content. Twitter policy requires developers to respect user intent by only republishing the most current version of a tweet or user profile in cases where such content has been modified, and developers must likewise perform timely deletion of content as needed (Twitter, Inc., 2017a, Developer Policy section 1.C.3). Respect for user intent does impact data sharing and republishing content, as discussed below, but by implication there is no policy requirement for obtaining user consent to analyze data collected via API. In contrast to other social media platforms such as Facebook and private chat or messaging services, Twitter's default public mode for tweets, the stated data use cases above, and the platform's comparatively broad privacy settings (Zimmer & Proferes, 2014a) may be reasonably understood to constrain user's expectations of privacy and consent and to support analytic use.

Two factors in particular argue against taking such a broad view. First, as noted by Golder et al. (2017), users who give their consent to public broadcast of social media may have context specific intentions and expectations about how the content they create will be used. Raymond (2010) documents, for example, the negative reactions of some users to the announcement of Twitter's donation of a complete data archive to the U.S. Library of Congress (Raymond, 2010, as cited in Zimmer & Proferes, 2014a). In a worst-case example, Crawford and Finn describe how social media data in the aftermath of a disaster can be used by well-meaning emergency responders in a way that disenfranchises vulnerable populations (Crawford & Finn, 2015). As with privacy, context is meaningful and there is a limited range of ethical uses for data that users share in a crisis (Crawford & Finn, 2015).

A second sticking point relates to relying upon terms of service (TOS) as mechanisms of consent. Specifically, the opacity of the terms themselves and the fact that few users bother to read them are issues addressed by Van Dijck (2013), Golder et al. (2017), and Luger, Moran, and Rodden (2013). In cases where users don't read TOS before implicitly accepting them through use of a platform, the ethical risk for researchers is that such implicit agreement may be a type of "uninformed consent" that would be considered unacceptable in some offline contexts.

In summary, the literature addressing social media and Twitter research ethics as described touches on complex questions of user agency and autonomy that are significant across research contexts. An overarching concern is the understanding or definition of social me-

dia users as research subjects and whether or where to make a distinction between people and data (Markham & Buchanan, 2012). Taking a data-centric view prioritizes researchers and third party data consumers over platform users. This can result, as noted, in the collection of large datasets with unaccounted sensitivities to vulnerable populations and violations of users' privacy. Conversely, a completely user-centric view is impractical and an impediment to research, as implied by the 2012 recommendations of the ethics working group of the Association of Internet Researchers (Markham & Buchanan, 2012). By recognizing the variety of social media platforms and the different degrees of direct and indirect user interactions afforded by different platforms, the guidelines acknowledge the impracticability of uniform, strictly policy-based approaches to ethical SMR conduct. Similarly, findings reported by Carter et al. (2015) document a variety of attitudes among researchers regarding the necessity and practicality of ethical review board participation across SMR contexts. For example, a blanket assertion that Twitter users are human subjects would consequently require the development and implementation of Institutional Review Board (IRB) and consent procedures for which there are currently few workable models (Carter et al., 2015).

The many nuances and context sensitivities of different SMR projects make it difficult therefore to define global best practices, as recognized and addressed by the 2012 recommendations of the AoIR Ethics Working Committee (Markham & Buchanan, 2012). Without being prescriptive, the recommendations provide a conceptual foundation for navigating ethics questions around SMR. In addition to encouraging researchers to consider the above noted distinction between people and data, the guidelines similarly address the evolving nature of human subjects and what counts as private versus public behavior.

Twitter's Changing Business Model: Research Implications of TOS and API Constraints

The expanding milieu of content creation across social and political contexts has increased the use of Twitter as a resource for research. Bruns and Burgess (2016) document the increasing scope of Twitter-based studies over time, while Zimmer and Proferes (2014b) elaborate not only on the breadth of Twitter research across disciplines, but also the evolving scope and scale of datasets (Zimmer & Proferes, 2014b). This growth in Twitter-based research has resulted in increased awareness of SMR issues related to data-collection methodologies, API policies, and differences in researchers' financial and technical resources. The literature addressing these issues is, to some extent, a literature of changing business models and their collateral impact on research.

Business models of social media platforms have necessarily changed over time to meet the

needs and expectations of users and advertisers. While research is an acknowledged use case per the current TOS (Twitter, Inc., 2017c), the needs and requirements of academic researchers likely have little impact on business models. Yet although academic Twitter research in particular has not historically driven platform improvements, it has at times spurred data access constraints and limitations. The most controversial changes have occurred since 2011, when, arguing that public archiving of datasets violated the API terms of service, Twitter ordered the shutdown of the popular Twitter research utility *TwapperKeeper*. This move was followed by a gradual tightening of API policies and implementation of data access restrictions that had a negative impact on research. Bruns and Burgess (2016) provide a detailed discussion of the circumstances surrounding these events.

Crucially, these changes brought about an end to informal practices that supported research. Prior to the update and from the platform's launch in 2006, Twitter's stance toward data access and development emphasized openness and transparency in support of community engagement to the benefit of the platform (Bucher, 2013; Van Dijck, 2013). This original openness was based on the expectation that third parties would be able to implement platform improvements and value additions that the founders had not envisioned (Bucher, 2013).

On the side of noncommercial research, the company's initial transparency was most apparent in the informal practice of whitelisting. To prevent abuse or resource constraints the API has always been subject to limitations. As detailed by Bruns and Burgess (2016), prior to 2011, researchers desiring privileged access to the API could request that they be whitelisted, or provided API access without restrictions. As the practice spread and researchers became more adept at interfacing with the API, Twitter datasets grew accordingly. The growth in the number and size of Twitter studies resulted in a corresponding proliferation of research platforms, which lowered the technical barriers and thus fostered additional research, etc. (Bruns and Burgess, 2016). With the tightening of its developer policies in 2011, Twitter subsequently ended the practice of whitelisting. The chilling effect of Twitter's changed posture toward market and noncommercial research is noted in greater detail by Bucher (2013) and Van Dijck (2013). Without claiming the TOS and API updates as the sole cause, Zimmer and Proferes (2014b) likewise note a decrease in the size and number of Twitter studies since 2011 (Zimmer & Proferes, 2014b, p. 257).

Consolidated control over data as a business asset is regarded as an important additional driver behind the policy and API changes (Burgess & Bruns, 2015; Puschmann & Burgess, 2013). With the discontinuation of whitelisting and newly imposed API limitations, access to unrestricted data became limited to Twitter's designated business partners, who in turn sell data at premium costs to third parties. The result is a privileging of data ac-

cess to companies and researchers with sufficient financial resources to pay premium data costs (Bruns & Burgess, 2016; Gaffney & Puschmann, 2014). Privileging users in this way necessarily privileges their corresponding research agendas. As reported by Bruns and Burgess (2016), some possibility of more equitable access to data was originally afforded by Twitter's "Data Grants" competition and the company's gift of a complete and ongoing archive of public tweets to the US Library of Congress (Bruns & Burgess, 2016; Raymond, 2010). However, the impact of both programs has proved minimal—few Data Grants have been awarded, and to date the Library of Congress has not been able to provide meaningful access to the archive. Additionally, on December 26, 2017, the Library of Congress announced that it will no longer collect every tweet, but will instead collect tweets "on a very selective basis" (U.S. Library of Congress, 2017, p. 1).

In addition to the ethical ramifications of Twitter's TOS and developer policies, the structure of the platform and API access points likewise impose constraints on research. While access to complete Twitter data is limited to designated business partners, every registered user is provided with default access to a stream of up to 1% of real-time data. Consequent problems of Twitter research relate to the representativeness of such data. Here, the nature of APIs and algorithms as non-neutral entities affecting data collection is significant (Van Dijck, 2013; Bucher, 2013; Puschmann & Burgess, 2013). First, the platform architecture shapes user interaction to promote certain users over others by filtering and weighting content (Van Dijck, 2013), which may ultimately bias a sample. Second, because APIs are non-neutral they likewise constrain the methodologies, tools, and types of research that can be performed through them (Borra & Reider, 2014; Bruns & Burgess, 2016; Burgess & Bruns, 2015). As noted by Puschmann and Burgess (2013, p. 45), TOS and APIs serve as "material instantiations of regulatory instruments used by the platform provider." With regard to Twitter, a particular manifestation of this regulation is via the different streaming APIs, which enable varying degrees of real-time access to tweets. In addition to the hard limits set by policy, the streaming API can further limit or block clients that lack sufficient bandwidth to quickly process received data. Twitter's concerns about data caching are legitimate, but this limitation means that without sufficient technical capacity or support, researchers may not even be able to effectively utilize the single streaming API to which all Twitter users have access.

Finally, even in scenarios where data collection is not constrained by technical limitations, several studies cite concerns over the completeness of datasets and how representative data is of all Twitter users who participate in a conversation or thread. As noted by Bonilla and Rosa (2015) and Lorentzen and Nolin (2017), data for hashtag studies, for example, may be incomplete as users will continue a thread or conversation but discontinue use of the relevant hashtag(s). A similar problem exists for studies of follower networks and user

mentions or retweets, when important or influential users may not be very active (Lorentzen, 2014, p. 330) or otherwise easily associated with a conversation (Lorentzen & Nolin, 2017). Lorentzen and Nolin (2017) present a method for accessing a greater percentage of data relevant to a given conversation, though their described procedures will be constrained by API limitations.

The above discussion about making ethical distinctions between people and data in SMR is also relevant to the research implications of Twitter's API and developer policies. From a primarily data-centric standpoint, API and policy constraints limit access to data with tremendous research potential to privileged clients capable of paying premium costs. From a more user-centric perspective, Twitter's data access controls promote data collection as a negotiation via API between researchers and Twitter that comes with pros and cons for users. On the one hand, the API formalizes respect for user intentions by returning only the most current versions of tweets and profile information. On the other, the API enables bulk collection and filtering of data for academic and market research use of which users may not be aware.

To conclude, available literature on Twitter and social media research illustrates how assumptions regarding appropriate and acceptable data use may be too easily made given the perceived public nature of much Twitter content. While not inherently harmful, these assumptions are problematic to the extent that they result from a data-centric view that prioritizes research interests over those of the individuals and communities represented within a dataset. With respect to Twitter, the data is further problematized by access controls imposed via policy and API constraints. While founded on legitimate business and technical concerns, limits to data access necessarily privilege the research agendas of organizations and institutions that can afford premium data access. For researchers without significant resources, the nature of publicly available data as a sample of a potentially much larger population begs questions of completeness and representativeness that likewise carry ethical implications.

DESCRIPTION OF THE PROGRAM

Data Collection: Realities of the API

Data for our study was gathered in phases using different methods as we developed our understanding of the complexities of Twitter data and as we increased to the extent possible our capacity to harvest, store, and analyze tweets and associated metadata. The data was collected using superseded versions of the TOS, developer agreement, and API. Where mean-

ingful, relevant differences with the current policies will be identified. Generally, the API access points and capabilities at the time of data collection were the same as those described by Gaffney and Puschmann (2014).

The description which follows does not go into extensive technical detail, but is presented rather as an overview of data collection as informed and impacted by Twitter's TOS, developer rules, and API. Referring to the methods available via the current Twitter API, data was collected using APIs for standard search (https://developer.twitter.com/en/docs/tweets/search/overview), timeline (https://developer.twitter.com/en/docs/tweets/timelines/overview), and streaming (https://developer.twitter.com/en/docs/tweets/filter-realtime/overview/statuses-filter). These are public APIs to which free access is granted to all registered Twitter users. Table 1 provides an overview of the API methods used, along with an example API request and information about current rate limits (as of May 2018).

API Method	Example request	Rate limits
Standard search	https://api.twitter.com/1.1/search/tweets.json?q=MAGA	450
User timeline	https://api.twitter.com/1.1/statuses/user_timeline.json?screen_name=realDonaldTrump	1500
Filter (streaming API)	https://stream.twitter.com/1.1/statuses/filter.json?track=MAGA	NA

Table 1. API methods used for data collection. Rate limits refer to the number of requests which can be made per fifteen minutes and are current as of May 30, 2018.

The standard search API enables querying for tweets that include specific terms or hashtags, user mentions, creation dates, etc. In addition to data gathered via the timeline and streaming APIs, querying the search API provided a complementary and focused method for collecting tweets relevant to specific topics or users. Use of this API for research purposes carries two important constraints. First, while a more extensive archive is available to premium and enterprise users at cost, results returned via the public search API are limited to the last seven days. This limitation did not affect us, however, as development of data-collection routines began in late November, 2016, well ahead of Trump's inauguration and our time frame of interest. As a result, there was little likelihood that we would need to search for tweets that were more than a week old.

Second, the standard search API is rate limited. That is, there is a limit to the number of times an individual user or application may execute a specific action within a given time frame. Rate limits are specific to different API methods.[2] Because our data collec-

[2] Current rate limit information for all HTTP methods is available at https://developer.twitter.com/en/docs/basics/rate-limits.html

tion utilized application authentication and the "GET search/tweets" method, we were limited to 450 search requests every fifteen minutes. We note that an individual "request" in this case can return multiple tweets, and it is therefore possible to stay within the rate limit and collect a sizable amount of data. The challenge for researchers lies in maximizing automated data collection without exceeding rate limits, as Twitter will block and may permanently ban users and applications that violate these limits.[3] Though our data collection tool was never permanently banned, it was periodically blocked for exceeding rate limits. This resulted in occasional gaps in data collected via the standard search API.

The "GET statuses/user_timeline" method of the timeline API returns a set of recent tweets from a specified user. From the standpoint of data collection, this amounts to automated "following" of that user. This method was used to execute periodic downloads of recent tweets from two users, @realDonaldTrump and @trumpRegrets.[4] Similar to the search API, timeline requests are rate limited, with the current limit of 1500 requests per 15 minutes exceeding the limit of 300 requests per 15 minutes in place at the time of our data collection. This limit did not negatively affect us since we only made requests every hour from @realDonaldTrump's timeline and every half hour from @trumpRegrets.

Finally, the streaming API was utilized as the best means of capturing large amounts of data. Unlike the search and timeline APIs, the public streaming API is not rated limited but instead enables collection of up to 1% of all tweets as they are posted in real time. Streaming API requests can be filtered, so it is possible when filters are applied to streaming API requests that the complete set of tweets meeting specified criteria will be collected if the total falls within the 1% limit (Burgess & Bruns, 2015).This detail bears some elaboration, as it directly affected our data collection. An unfiltered request to the public, "free" version of the streaming API will at most return 1% of all tweets published in the moment when that request is made. If a thousand tweets are published in that moment, an unfiltered streaming API request will return 1%, or ten tweets. Filtering allows more targeted data collection within the 1% cap. For example, filtering streaming API data for a specific hashtag such as #MAGA means that if the hashtag is present in fewer than 1% of all tweets published at the moment of the request, then every tweet meeting the filter

[3] Libraries including Python Tweepy (http://www.tweepy.org/) provide methods that automatically accommodate Twitter's rate limits. Though we ultimately made use of Tweepy to collect data from the streaming API, we preferred to develop our own processes for collecting data from the search and user timeline APIs.

[4] While our original intent was only to follow @realDonaldTrump, the emergence of this user profile ahead of the inauguration presented a potential opportunity to capture alternative or ironic uses of the #MAGA hashtag.

condition will be returned. That is, it is possible at times to get "all" relevant data via the public streaming API. However, when a streaming API request is filtered on a trending hashtag, it's possible that more than 1% of all tweets published in a given moment will contain that hashtag. In such cases, the returned data will be incomplete because relevant tweets will be excluded. As noted below, this limit had a definite impact on our streaming API data collection.

The public version of the streaming API has an additional constraint relevant to data collection for research purposes. Only one streaming request is allowed per registered Twitter user, so whereas we had multiple concurrent requests collecting data from the search and timeline APIs, we had to combine all of our filter criteria into a single streaming API request. Using the Python Tweepy library (Tweepy, 2015, version 3.6.0), we were able to implement the request as described in Table 2.

For requests made via the streaming API, multiple filter terms are combined using an "or" operator. Thus, the tweets returned via our streaming API request could meet any of the specified filtering conditions. Once the total number of matching tweets exceeded the 1% limit, we had no control over which tweets were excluded from the streamed data and were at the mercy of the underlying algorithms. The streaming API provides information about how many tweets have been excluded from a request once the 1% limit is exceeded, so it is possible to gauge the volume of missed data.

Utilizing the streaming API resulted in robust data collection, though was subject to interruptions and outages due to routine system upgrades, power outages, and unexpected bugs. Consequently there are gaps in the streaming data. In particular, maintaining a persistent connection sufficient for real-time data capture proved difficult. Troubleshooting this problem delayed data collection from the streaming API, with the result that we began capturing this data after the inauguration on January 26, 2017. Thus, in addition to gaps from outages we have a gap immediately following the inauguration, during the eventful first week of the administration. We note that we continued to run our search and user timeline harvest utilities as complementary processes, so we have some data for the entire period between late November 2016 and May 2017.

Table 2 provides an overview of the different API methods and query or filter terms which were used to collect data, the date range in which data was collected per API method, and some summary information about the final dataset composition. Data collection methods via search and timeline APIs resulted in downloading multiple copies of particular tweets, so deduplicated data is provided in Table 2.

API method	Query or filter terms	Collection date range	Total unique tweets	Total unique users
Standard search	@realDonaldTrump user mention	11/28/16 - 5/1/17	4,597,326	1,501,806
Standard search	"Trump" in text	01/18/17 – 5/1/17	11,055,772	2,648,849
Standard search	#MAGA hashtag	01/23/17 – 5/1/17	1,169,897	236,033
User timeline	realDonaldTrump	12/21/16 – 5/1/17	902	1*
User timeline	trumpRegrets	01/15/17 – 5/1/17	1,751	1*
Filter (streaming API)	Follow realDonaldTrump; Search for #maga hashtag, @realDonaldTrump user mention, "Trump," or "POTUS" in text	01/26/17 – 5/1/17	201,447,504	12,489,255

Table 1. Overview of data collected via different public Twitter APIs. *As the user timeline API only returns tweets from the specified user, the count of one unique user within timeline data may not be illustrative of the variety of users whose tweets are quoted or retweeted within a dataset.

Implications and Lessons Learned

The author's experience with data collection via the Twitter API illustrates multiple issues described in the literature, including the complexities of SMR ethics and the effects which corporate control of social media data can have on research. Implications for scholarly communications librarianship relate to emerging practices in data literacy and data management, with particular regard to big data ethics, research reproducibility, and the sharing and archiving of social media datasets.

The most significant lessons learned relate to the many nuances of privacy and ethics within Twitter research. In particular, the definition of what counts as a human subject within SMR is context sensitive and available IRB guidance may not adequately address this ambiguity. Our research provides a case in point: Prior to collecting data, we did not directly consult with IRB staff but instead referred to a decision tree published by the IRB to determine whether our study would be considered human subjects research (HSR) (University of New Mexico Institutional Review Board, 2015). Because we intended to collect existing data that was publicly available, per the decision tree our research did not

meet the definition of HSR and as such did not require IRB review.[5]

However, by effectively "self-exempting" our study from IRB review, we sidestepped issues of agency and privacy which as discussed have become contested within the context of social media publics and which suggest a need to revisit definitions of HSR. The self-exempting rationale constitutes a researcher rather than user or human subject–oriented focus that undermines the agency and intent of Twitter users. For example, our initial question in working through our IRB's decision tree was the definition of "existing data" and whether or to what extent our API data collection constituted any kind of user interaction. Having through the course of our research developed a more nuanced understanding of the ethical grey areas in SMR, the definitional question has since become what counts as "public." Specifically, as noted in the literature, a user's intention to tweet cannot be taken as an intention to broadcast information publicly. Our concern in this regard is not only with tweet text but also and to a larger degree with other default public information including user profiles. The question bears further research, but while many if not most Twitter users may accept the public nature of their status updates, we question the level of user awareness and acceptance regarding the public nature of their profile information. Recalling Mao et al. (2011), a similar study of privacy leaks in profile information is justified.

Regardless of whether librarians interact directly with Twitter or social media data as researchers, implications of Twitter and SMR data ethics for scholarly communications librarianship are twofold. First, librarians provide information and training about data ethics through data literacy initiatives and participation in institutional responsible conduct of research programs. Given the ongoing popularity of Twitter studies and SMR in general, the evolving definitions and context sensitivities of privacy and consent within SMR are important issues to address with the students and new researchers who typically attend such sessions. A second and similar implication exists for data services units or any librarians that provide regular data management or data management planning consultations to researchers. Data management planning guidelines from the NSF (National Science Foundation, 2017, chapter II section C2j), among others, specifically require researchers to address data sensitivities including privacy. Through increased awareness of these issues, the identification of SMR data sensitivities and mitigation strategies can enhance the impact of DMP services.

[5] The author notes that as a result of our research into SMR ethics we did contact and consult directly with our IRB prior to beginning content analysis of the data. The IRB has confirmed our original determination that no review is or was required.

The impact of corporate control of data with high research value is a second area in which our study demonstrates Twitter and SMR issues described in the literature. As noted in the literature review, Twitter's stance toward researcher access to data has changed and generally become more restrictive over time. Currently, access to the complete Twitter database is limited to a handful of partner companies which in turn sell the data at a cost few public or academic researchers can afford. Aside from privileging the research agendas of corporations and whichever academic researchers can afford these costs, further issues include the completeness and unknown biases present within data samples collected via the more limited but freely available public APIs.

Two details regarding our streaming data illustrate these issues. First, as noted in Table 2, whereas the data collected via the streaming API consists of 201,447,504 unique tweets, the number of unique users who posted those tweets is only 12,489,255. These numbers indicate that subsets of users may be under- or overrepresented within the dataset, but to what extent this results from actual user behavior or is an artifact of the streaming API-filtering algorithms requires investigation. Second, using the information provided via the streaming API about missed tweets, or the number that exceeded our 1% allotment, we've been able to determine that we missed a total of 706,046,014 tweets. This difference is a result of the limits imposed by Twitter on the public streaming API. Both details are problematic and available strategies for us to address any corresponding impact on our sample will be discussed below. We include these statistics here, however, as an illustration of the effect of Twitter API policies on SMR and as supporting evidence of Burgess and Bruns's (2015) observation that Twitter data collected via the APIs carries implications regarding representativeness and the characteristics of the data as a sample of a larger population (Burgess & Bruns, 2012; Burgess & Bruns, 2015).

Implications for scholarly communications librarianship in terms of data literacy and data management support are similar to those noted above regarding privacy and research ethics. The effects of data access limitations and biases within samples are issues that librarians can describe and explore with undergraduate and graduate students using literature and examples from social media research. Especially important to note in this regard are the limitations of reproducibility. Considering that Twitter users may edit or delete tweets or even delete their accounts, datasets shared according to Twitter policy as described below are at best snapshots, which may not be completely reproduced by other researchers. This further affects data sharing and archiving services if and when library-supported institutional and data repositories are used to host Twitter datasets. Twitter policy prohibits sharing full datasets, requiring researchers to instead publish more limited sets of information including individual tweet and user identifiers. Librarians who manage repositories need to be aware not only of current Twitter policy but also methods

and tools for extracting and publishing identifiers that satisfy policy requirements.

NEXT STEPS

As we proceed with content analysis, next steps include defining and implementing methods for mitigating the ethical and technical issues addressed in this paper.

API restrictions continue to impact our research after data collection. Gaps in the data are difficult to fill without paying for access to the search archive. Thus, a key question to address through initial analysis is the representativeness of the dataset. Completeness is a known issue as indicated by Table 2 above. Beyond that, as in many Twitter studies, the scope of our data collection was defined primarily through hashtags and user timelines. Completeness is therefore more than just an issue of missed tweets because, as noted, users may stop using a hashtag but continue a relevant conversation. Similarly, with regard to the impact of retweeting behavior and the development of follower or mention networks, different retweeting and user mention practices across a large sample mean that retweets may not always be identified as such and influential users may not be well represented either in the dataset or aggregate analysis.

Lorentzen and Nolin (2017) describe a method for using the API to retrieve tweets not included within a dataset but which are replied to by tweets within the dataset. For the study described, the authors note improved retrieval of users and tweets relevant to a Twitter thread. Our dataset is much larger than was used in Lorentzen and Nolin's study—we expect rate limits will constrain our ability to collect additional data, but we are investigating the application of their method to our dataset.

As a matter of respecting user intent and privacy, while we may analyze individual tweets as parts of an overall conversation, results will be reported in the aggregate and we will not republish individual user information or statuses. First, it is possible that our dataset contains nonpublic information such as direct messages or private tweets that were inadvertently made public through retweeting (Mao et al., 2011). Second, remaining sensitive to privacy and consent within a default public system like Twitter is a matter of respecting user intent to the greatest extent possible. While that may seem vague, some guidance is provided by the Developer Policy (Twitter, Inc., 2017a) section 1.C, "Respect User's Control and Privacy." In part this section was discussed above regarding developer activities that require user consent. Additional relevant clauses prohibit the reidentification of persons, households, or other user characteristics using anything other than public information (section 1.C.2), and content modifications (section 1.C.3). Specifically, if a tweet is deleted or modified, or if a user's profile is deleted or updated, a reasonable effort should be made to update the dataset

to include that modification. This clause has an implication for sharing datasets that will be addressed below.

We acknowledge that the nature of our research and our primary research questions require us to highlight the activity of a particular individual, notably the president of the United States, via his @realDonaldTrump account, and, to a lesser extent, the @POTUS account. However, status updates via the @POTUS account are in the public domain, and there is precedent for making the same case for tweets from his personal account as well, as long as he holds the office of president. That said, any @realDonaldTrump or @POTUS tweets we may eventually seek to republish will be checked for deletion or modifications. In keeping with Twitter policy, deleted tweets will not be republished and modified tweets will be presented as updated or modified.

Finally, while we have been at times unsure about whether we may or may not publish our dataset, the current version of the developer policy provides for academic or noncommercial data sharing. Specifically, whereas the Developer Policy (Twitter, Inc., 2017a, section 1.F.2) places limits on providing content to third parties via direct download, an explicit exemption is made for sharing "on behalf of an academic institution and for the sole purpose of non-commercial research" (Twitter, Inc., 2017a, section 1.F.2.b).

Here a distinction between tweet identifiers and the tweet object can be made, which illustrates the point above about honoring user intent and respecting potential modifications to content. Tweets as objects contain not just the tweeted text or status update, but additionally contain the corresponding author's profile information and, in the case of a retweet, the profile information of the user who originally posted the tweet. Much of this information is subject to modification, but only the most current version of a user's profile or a tweet that hasn't been deleted can be returned from the Twitter API using the tweet identifier. If a tweet or a user has been deleted, the API will in such cases return an error or empty message. Thus, so long as tweet datasets are shared in accordance with the policy restriction to publish only tweet and user identifiers, a harvest of those identifiers will only return the most up-to-date content.

As a next step, therefore, we are currently preparing a dataset of tweet indentifiers for sharing. However, policy and data management processes aside, there are larger issues to address before making the dataset public. Specifically, our dataset likely contains expressions of political opinions that are unpopular now or may be so in the future, in addition to potentially controversial statements of belief regarding gender, sexuality, or any of a number of culturally sensitive topics. Deletions or modifications to potentially controversial tweets are accounted for by Twitter's dataset sharing policy as described above, but we find it unlikely

that many or most sensitive tweets are ever modified or deleted.

Even so, compelling reasons exist for sharing the dataset. First, reproducibility of research can only be supported if the data is made public. Some publishers require data sharing, and in any case as librarians and researchers this is a principle we support and for which we advocate. Second, sharing Twitter datasets democratizes access to the Twitter index that is otherwise only available to corporations and research institutions with significant resources and deep pockets. Shared datasets provide a means for retrieving historical data that would only otherwise be available at cost.

Before publishing our dataset we will therefore continue to evaluate privacy issues around sharing Twitter data, especially as sharing regardless of rationale is inherently data-centric and any accommodations we can make to protect privacy are likely founded on perceptions of our own risk rather than user agency. For example, whereas Twitter policy allows sharing of user as well as tweet identifiers, a review of datasets within the Documenting the Now[6] registry indicates that inclusion of user identifiers is uncommon. This is a reasonable safeguard, which to some extent shields users from untoward surveillance. Using our own dataset as an example, we note from Table 2 that our collection of data from the streaming API consists of over 200 million tweets but only 12.5 million users. If we share tweet identifiers with their corresponding user identifiers, filtering out the most active users becomes a trivial exercise that can be accomplished in minutes using any database application. Anyone with this data could then execute a targeted harvest of recent tweets and profile information for that subset of users. If we publish only tweet identifiers, it is still possible to obtain this information but not without "hydrating" a significant portion of the dataset. A description of the requirements for hydrating a set of tweet identifiers is beyond the scope of this discussion, but the process is complex enough to offer some deterrent to casual abuse of tweet identifier datasets.[7] It is not, however, a complete deterrent.

CONCLUSION

Among social media platforms, Twitter exemplifies the types of information sharing and community coherence that can blur distinctions between public and private behavior and add complexity to understandings of human subjects and consent in SMR. Definitions are fluid and ultimately depend on the context of specific studies. To this extent, we have presented our data collection as a case study with the purpose of illustrating these issues in

[6] http://www.docnow.io/

[7] An overview of the hydration process is provided by Ed Summers at https://medium.com/on-archivy/on-forgetting-e01a2b95272.

practice, along with corresponding problems of corporate constraints on data access. As mentioned, there are many implications for scholarly communications librarianship related to SMR ethics and data access policies. Significant areas include democratizing access to data by advocating for and participating in ethical and policy-based sharing, as well as proactive engagement with vulnerable populations to provide information and resources about responsible social media use, along with information about how and by whom data may be accessed. As SMR continues to grow, how and to what extent scholarly communication and research librarians engage with these issues is an area that requires further research. It is our hope by that surfacing the complexities surrounding our Twitter data collection, our experience may provide some perspective on these issues for other librarians and researchers working with social media data.

ACKNOWLEDGMENTS

The author would like to acknowledge Dr. Teresa Neely, assessment librarian at the University of New Mexico College of Libraries and Learning Sciences, for her significant contributions to an earlier version of this article and the ongoing related research.

REFERENCES

Bonilla, Y., & Rosa, J. (2015). #Ferguson: Digital protest, hashtag ethnography, and the racial politics of social media in the United States: #Ferguson. *American Ethnologist*, *42*(1), 4–17. https://doi.org/10.1111/amet.12112

Borra, E., & Rieder, B. (2014). Programmed method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278. https://doi.org/10.1108/AJIM-09-2013-0094

Boyd, Danah. (n.d.). Bibliography of Research on Twitter & Microblogging. Retrieved from https://www.danah.org/researchBibs/twitter.php

Bruns, A., & Burgess, J. (2016). Methodological innovation in precarious spaces: The case of Twitter. In H. Snee, Y. Morey, S. Roberts, & H. Watson (Eds.) *Digital Methods for Social Science: An Interdisciplinary Guide to Research Innovation* (pp. 17–33). London: Palgrave MacMillan. https://doi.org/10.1057/9781137453662_2

Bucher, T. (2013). Objects of intense feeling: The case of the Twitter APIs. *Computational Culture*, *3*. Retrieved from http://computationalculture.net/objects-of-intense-feeling-the-case-of-the-twitter-api/

Burgess, J., & Bruns, A. (2012). Twitter archives and the challenges of "Big Social Data" for media and communication research. *M/C Journal*, *15*(5). Retrieved from http://journal.media-culture.org.au/index.php/mcjournal/article/view/561

Burgess, J., & Bruns, A. (2015). Easy data, hard data: The politics and pragmatics of Twitter research after the computational turn. In G. Elmer, G. Langlois, & J. Redden (Eds.), *Compromised Data: From Social Media to Big Data* (pp. 93–111). London: Bloomsbury.

Carter, C. J., Koene, A., Perez, E., Statache, R., Adolphs, S., O'Malley, . . . McAuley, D. (2015). Understanding academic attitudes towards the ethical challenges posed by social media research. *ACM SIGCAS Computers and Society*, *45*(3), 202–210. https://doi.org/10.1145/2874239.2874268

Crawford, K., & Finn, M. (2015). The limits of crisis data: analytical and ethical challenges of using social and mobile data to understand disasters. *GeoJournal*, *80*(4), 491–502. https://doi.org/10.1007/s10708-014-9597-z

Gaffney, D., & Puschmann, C. (2014). Data collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 55–68). New York: Peter Lang.

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. https://doi.org/10.2196/jmir.7082

Kalsnes, B., Krumsvik, A. H., & Storsul, T. (2014). Social media as a political backchannel: Twitter use during televised election debates in Norway. *Aslib Journal of Information Management*, *66*(3), 313–328. https://doi.org/10.1108/AJIM-09-2013-0093

Lorentzen, D. G. (2014). Polarisation in political Twitter conversations. *Aslib Journal of Information Management*, *66*(3), 329–341. https://doi.org/10.1108/AJIM-09-2013-0086

Lorentzen, D. G., & Nolin, J. (2017). Approaching completeness: Capturing a hashtagged Twitter conversation and its follow-on conversation. *Social Science Computer Review*, *35*(2), 277–286. https://doi.org/10.1177/0894439315607018

Luger, E., Moran, S., & Rodden, T. (2013). Consent for all: Revealing the hidden complexity of terms and conditions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2687–2696). New York: ACM. https://doi.org/10.1145/2470654.2481371

Mao, H., Shuai, X., & Kapadia, A. (2011). Loose tweets: An analysis of privacy leaks on Twitter. In *Proceedings of the 10th annual ACM Workshop on Privacy in the Electronic Society* (pp. 1–12). New York: ACM. https://doi.org/10.1145/2046556.2046558

Markham, A., & Buchanan, E. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (version 2.0)*. Retrieved from https://aoir.org/reports/ethics2.pdf

National Science Foundation (2017, January 30). Proposal & award policies & procedures guide. Retrieved from https://www.nsf.gov/pubs/policydocs/pappg17_1/pappg_2.jsp#IIC2j.

Puschmann, C., & Burgess, J. (2014). The politics of Twitter data. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.) *Twitter and society* (pp. 43–54). New York: Peter Lang.

Raymond, M. (2010, April 14). How tweet it is!: Library acquires entire Twitter archive [Blog post]. Retrieved from http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/

Risse, T., Peters, W., Senellart, P., & Maynard, D. (2014). Documenting contemporary society by preserving relevant information from Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.) *Twitter and society* (pp. 207–219). New York: Peter Lang.

Robbins, M. L. (2017). Practical suggestions for legal and ethical concerns with social environment sampling methods. *Social Psychological and Personality Science*, *8*(5), 573–580. https://doi.org/10.1177/1948550617699253

Schmidt, J.-H. (2014). Twitter and the rise of personal publics. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 3–14). New York: Peter Lang.

Summers, Ed. (2014, November 18). On forgetting and hydration. [Blog post]. Retrieved from https://medium.com/on-archivy/on-forgetting-e01a2b95272

Taylor, J., & Pagliari, C. (2017, October 26). Mining social media data: How are research sponsors and researchers addressing the ethical challenges? *Research Ethics*, 14(2), 1-39. https://doi.org/10.1177/1747016117738559

Tweepy [computer software], v3.6.0. (2015). Retrieved from https://github.com/tweepy/tweepy

Twitter, Inc. (n.d.). Developer rules of the road. Retrieved from https://dev.twit-ter.com/terms/api-terms

Twitter, Inc. (2017a, November 3). Developer agreement and policy. Retrieved from https://developer.twitter.com/en/developer-terms/agreement-and-policy

Twitter, Inc. (2017b, June 18). Twitter privacy policy. Retrieved from https://twitter.com/en/privacy

Twitter, Inc. (2017c, October 2). Twitter terms of service. Retrieved from https://twitter.com/en/tos

Twitter, Inc. (2018). API reference index. Retrieved from https://developer.twitter.com/en/docs/api-reference-index

United States Library of Congress. (2010, April 14). Gift agreement. Retrieved from https://blogs.loc.gov/loc/files/2010/04/LOC-Twitter.pdf

United States Library of Congress. (2017, December). Update on the Twitter archive at the Library of Congress. Retrieved from https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf

University of New Mexico Institutional Review Board (2015, September 14). Does your project require UNM IRB review? Retrieved from http://irb.unm.edu/sites/default/files/Decision%20Trees_Does%20Project%20Need%20IRB%20Review.pdf

Van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199970773.001.0001

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C., Eds. (2014). *Twitter and society*. New York: Peter Lang. https://doi.org/10.3726/978-1-4539-1170-9

Zimmer, M., & Proferes, N. (2014a). Privacy on Twitter, Twitter on privacy. In Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (Eds.) *Twitter and society* (pp. 169–82). New York: Peter Lang.

Zimmer, M., & Proferes, N. J. (2014b). A topology of Twitter research: disciplines, methods, and ethics. *Aslib Journal of Information Management*, *66*(3), 250–261. https://doi.org/10.1108/AJIM-09-2013-0083