

# Archiving Scholars' Tweets

KAITLIN L. COSTELLO and JASON PRIEM

School of Information and Library Science, University of North Carolina at Chapel Hill

## Abstract

As a rich, personal source of data on a uniquely large and diverse set of users, Twitter merits long-term preservation – as the Library of Congress recognized in announcing the creation of a Twitter archive. We conducted 28 semi-structured interviews and used conventional qualitative content analysis to explore the attitudes and expectations that one sub-group of users – scholars – had towards having their tweets archived, and who they thought could be responsible for such an archive. Our participants questioned the utility of a comprehensive archive and were concerned by the possibility of their institutional employers archiving their tweets. Since the expectations of content creators can help to inform archival policy and practice, this research is valuable to the custodians of a Twitter archive, and may also be applied to other social media archives.

## Problem Statement

Twitter is a fast-growing web service which allows people to post messages, or “tweets,” of 140 characters or fewer online. Twitter is a rich, personal source of data on a set of users that is uniquely large (145 million registered users), diverse, and increasingly international (60% from outside the United States).<sup>1</sup> As such, some or all of these users’ tweets may merit long-term preservation. Certainly, in recent months, there has been much scholarly and popular discussion of such preservation.<sup>2</sup> This conversation has been largely spurred by the Library of Congress’ (LoC) announcement of a plan to establish a Twitter archive, although many of the details of what is meant by “archive” remain unclear. They have explained that the archive will not contain private tweets, will be long-term, and will only be accessible in its entirety to researchers upon request, although special collections may be publicly displayed.<sup>3</sup>

---

<sup>1</sup> Evan Williams, “The Evolving Ecosystem,” *Twitter blog*, September 2, 2010, <http://blog.twitter.com/2010/09/evolving-ecosystem.html>; Matt Sanford, “Growing Around the World,” *Twitter blog*, April 8, 2010, <http://blog.twitter.com/2010/04/growing-around-world.html>.

<sup>2</sup> Randall Stross, “A Sea of History: Twitter at the Library of Congress,” *The New York Times*, May 1, 2010, sec. Business, [http://www.nytimes.com/2010/05/02/business/02digi.htm?\\_r=1](http://www.nytimes.com/2010/05/02/business/02digi.htm?_r=1); Laura E. Campbell and Beth Dulabahn, “Digital preservation: The Twitter archives and NDIIPP,” in (presented at the iPres 2010: 7th International Conference on Preservation of Digital Objects, Vienna, Austria, 2010), <http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/campbell-27.pdf>.

<sup>3</sup> Matt Raymond, “How tweet it is!: Library of Congress acquires entire Twitter archive,” *Library of Congress Blog*, April 14, 2010, <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>; Campbell and Dulabahn, “Digital preservation: The Twitter archives and NDIIPP.”



Despite the growing excitement around archiving tweets, not all reactions have been positive. Some have expressed concern about whether a Twitter archive would violate users' expectations of privacy.<sup>4</sup> Others have complained that such an archive would be "a waste of space that could be better used for really historic books and data."<sup>5</sup>

However, although people have speculated about users' concerns over a Twitter archive, no one has systematically asked content creators themselves how they feel about having their tweets archived. This is an omission that should be corrected, because the expectations of content creators (especially those still living) have legal, ethical, and practical repercussions in archival practice.<sup>6</sup>

The goal of this study was to gain a better understanding of how content creators viewed the potential archiving of their tweets. Specifically, we wanted to know:

- R1.** Which of their tweets, if any, do content creators think should be saved and made available for future use?
- R2.** If tweets are to be archived, who do content creators think should be responsible for gathering and maintaining the collection of stored tweets?

## Methods

To answer these questions, we interviewed 28 Twitter users. As is common in Twitter research,<sup>7</sup> we chose to focus our study on one sub-group: scholars. In addition to using Twitter personally, scholars also use Twitter professionally: they cite peer-reviewed articles, "live tweet" notes at

---

<sup>4</sup> Fred Stutzman, "Twitter and the Library of Congress," April 14, 2010, <http://fstutzman.com/2010/04/14/twitter-and-the-library-of-congress/>.

<sup>5</sup> Dale McNamee, "Comment on How Tweet It Is!: Library Acquires Entire Twitter Archive," *Library of Congress Blog*, April 19, 2010, <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/#comment-35401>.

<sup>6</sup> Michael J. Crawford, "Copyright, Unpublished Manuscript Records, and the Archivist," *American Archivist* (1983): 135–147; Jeffrey Zeldman, "You cannot copyright a tweet, particularly the comments section," *The Daily Report*, February 25, 2010, <http://www.zeldman.com/2010/02/25/you-cannot-copyright-a-tweet/>; Society of American Archivists Council, "Code of Ethics for Archivists," February 5, 2005, [http://www.archivists.org/governance/handbook/app\\_ethics.asp](http://www.archivists.org/governance/handbook/app_ethics.asp); Richard J. Cox and Helen Willa Samuels, "The archivist's first responsibility: A research agenda to improve the identification and retention of records of enduring value," *The American Archivist* (1988): 28–42.

<sup>7</sup> Jennifer Golbeck, Justin M. Grimes, and Anthony Rogers, "Twitter use by the U.S. Congress," *Journal of the American Society for Information Science and Technology* 61, no. 8 (August 2010): 1612-1621; Thomas Heverin and Lisl Zach, "Twitter for City Police Department Information Sharing," in *Proceedings of the 73rd ASIS&T Annual Meeting* (presented at the ASIS&T 2010, Pittsburgh, PA, 2010); Jason Priem and Kaitlin Costello, "How and why scholars cite on Twitter," in *Proceedings of the 73rd ASIS&T Annual Meeting* (presented at the ASIS&T 2010 Annual Meeting, Pittsburgh, PA, 2010); Dejin Zhao and Mary Beth Rosson, "How and why people Twitter: the role that micro-blogging plays in informal communication at work," in *Proceedings of the ACM 2009 international conference on Supporting group work* (Sanibel Island, Florida, USA: ACM, 2009), 243-252, <http://portal.acm.org/citation.cfm?id=1531710#CIT>.

conferences, publicize research, and facilitate classroom discussions on Twitter.<sup>8</sup> When we began our study, the LoC announcement had not yet been made. We suspected that collections of scholarly tweets housed in institutional repositories could be the first type of Twitter archives, given scholars' professional use of Twitter and the value that scholarly culture places on the longevity of the written word.

In light of the announcement by the LoC, we believe our research is still relevant for two reasons. First, the LoC announcement split our sample neatly in half: fourteen interviews were completed before the announcement, and fourteen after. This gave us the unique opportunity to observe perspectives about archiving tweets both before and after the announcement. Second, although it is narrowly focused, it is the only systematic investigation into the opinions of content creators on archiving tweets to date. Because of its limited scope and qualitative nature, this research must be considered exploratory. Although it cannot decisively answer our research questions, it presents a good starting point for broader, more quantitative future efforts.

We used *conventional qualitative content analysis (CQCA)*<sup>9</sup> to examine our research questions. This and related qualitative methods have long been heavily employed in social sciences<sup>10</sup> because they allow researchers to “understand the social reality in a subjective but scientific manner.”<sup>11</sup> In this method, a set of purposively selected texts (in our case, the transcripts of semi-structured interviews) are analyzed to uncover themes in “expressions from subjects reflecting how they view the social world.”<sup>12</sup> In simpler terms: we asked scholars questions, then found the themes in their answers.

Our first step was to gather a purposive sample of 28 scholars, defined as people who said they were faculty members, post-docs, or doctoral students who reported using Twitter at least once a week in pre-interview screening. To contact a diverse group of scholars on Twitter, we used a snowball sampling method, beginning with a seed sample of three individuals: one scholar in the humanities, one in the social sciences, and one natural scientist. Each was selected because he had many dozens of academic followers. We asked them to tweet the following message:

*Are you an academic on Twitter willing to participate in a short research study? Full details and contact: <http://bit.ly/cVXYNs>.*

The link resolved to a recruitment webpage (Appendix A). Throughout the recruitment phase, we asked each new interviewee to tweet the message to his or her followers, gradually increasing the

---

<sup>8</sup> Priem and Costello, “How and why scholars cite on Twitter”; Julie Letierce et al., “Understanding how Twitter is used to spread scientific messages,” in (presented at the Web Science Conference 2010, Raleigh, North Carolina, USA, 2010); Jeffrey R. Young, “10 High Fliers on Twitter,” *The Chronicle of Higher Education*, April 10, 2009; Joanna C. Dunlap and Patrick R. Lowenthal, “Tweeting the night away: Using Twitter to enhance social presence,” *Journal of Information Systems Education* 20, no. 2 (2009).

<sup>9</sup> Yan Zhang and Barbara M. Wildemuth, “Qualitative analysis of content,” in *Applications of social research methods to questions in information and library science* (Westport, Connecticut: Libraries Unlimited, 2009), 308-319.

<sup>10</sup> Linda Schamber, “Time-line interviews and inductive content analysis: their effectiveness for exploring cognitive behaviors,” *Journal of the American Society for Information Science* 51, no. 8 (2000): 734–744.

<sup>11</sup> Zhang and Wildemuth, “Qualitative analysis of content,” 308.

<sup>12</sup> *Ibid.*, 309.

size of our potential sample. Our final sample of 28 participants contained 7 scientists, 14 social scientists, and 7 humanists, as determined by the academic department with which they were most closely affiliated.

Next, we conducted 30- to 45-minute long semi-structured phone interviews using an interview guide (Appendix B). The interview guide contained questions for both this study and a related study on scholarly impact on Twitter. Semi-structured interview guides are not inflexible scripts, but rather cover general conversation points.<sup>13</sup> After all of the interviews were complete, they were transcribed in full (8 by the researchers, and 20 by a professional transcriptionist) and pseudonyms were randomly assigned to the interviewees. We then used open coding to inductively categorize participants' responses; in this method, "themes and categories emerge from the data through the researcher's careful examination and constant comparison."<sup>14</sup> The resulting 31-page codebook contained 5 major themes: Properties of Twitter, Identity, Academics on Twitter, Saving Tweets, and Blogging & Microblogging. These were further divided into 35 sub-themes, supported by 315 quotations.<sup>15</sup>

Naturalistic methods, including CQCA, are not intended for analysis with statistical tools: "there is no statistical test of significance to determine if [qualitative] results 'count.'"<sup>16</sup> In fact, we have purposefully omitted counts or percentages for our qualitative findings (e.g., the number of scholars who supported saving all tweets) to avoid giving the impression that our data statistically sample a population in a generalizable way. This lack of quantifiability is a weakness of CQCA; however, this method also "addresses some of the weaknesses of the quantitative approach," allowing researchers to inductively build understandings from participants' unique and idiosyncratic responses, rather than fitting responses into pre-built, deductive structures.<sup>17</sup>

To augment our qualitative findings, we harvested the most recent 3,200 tweets – the maximum Twitter allows – from each of our 26 interviewees with public Twitter feeds. These were stored in a MySQL database and their text analyzed by counting rows returned from SQL regular expression searches (see Table 1).

## Results

### ***R1. Which of their tweets, if any, do content creators think should be saved and made available for future use?***

Responses to questions about which tweets should be "saved for long-term use" were mixed, with some participants wholly opposed to saving any tweets at all and some advocating that all tweets should be saved. We purposefully avoided technical definitions of "archive," "saving," "long-term," and "lasting value" in the interviews, because we wanted to access participants'

---

<sup>13</sup> Steinar Kvale and Svend Brinkmann, *InterViews: Learning the Craft of Qualitative Research Interviewing*, 2nd ed. (Sage Publications, Inc, 2009).

<sup>14</sup> Zhang and Wildemuth, "Qualitative analysis of content," 309.

<sup>15</sup> The complete codebook is available upon request to [kaitcost@email.unc.edu](mailto:kaitcost@email.unc.edu).

<sup>16</sup> Elliot W. Eisner, *The enlightened eye: Qualitative inquiry and the enhancement of educational practice* (Macmillan Pub. Co., New York, NY, 1991), 39.

<sup>17</sup> Zhang and Wildemuth, "Qualitative analysis of content," 308.

own understandings of these concepts; consequently, in our analysis, these terms must be understood broadly.

A majority of our participants said that only some tweets should be saved. Given the semi-structured nature of the interviews and the complexity of the topic, responses often varied throughout the course of each interview. However, while the categories of “none, some, or all” reflect a necessary oversimplification of the responses we received, they do capture the essence of participants’ views. Since this was qualitative research, we did not quantify these results, instead focusing on the substance of individual scholars’ voices.

### *None*

Some of our participants were opposed to having their tweets archived in any form, largely because they viewed content posted on Twitter as trivial. As Tameka said, “As far as I’m concerned, from a utility point of view, I don’t see much value in saving old tweets – anybody’s old tweets.” Several participants were drawn to the medium because of its ephemeral nature. Ronnie explicitly contrasted this ephemerality with the enduring nature of more formal scholarly writing:

**Ronnie:** [With peer-reviewed work], you’re writing something that you want to have meaning ten, fifty, 100, 1000 years from now. When you’re tweeting, it’s more just in the moment. I mean, you only have 140 characters, and you are sort of spitting it out there.

Participants frequently discussed their concern for present-day privacy and identity management:

**Elaine:** I’m planning on applying for an academic job, and if they go back and they look at my tweets and they find “Oh wow, she’s a raging Marxist,” or something, or “It sounds like she gets drunk a lot,” or, “She has an extensive doll collection, that’s creepy.” You know, these kinds of things can come back and influence you.

Several participants also indicated that the most worrisome type of Twitter archive would be one that was available while they were still alive, as the management of their archived identity would be out of their control. Derrick said, “This Library of Congress [archive] is kind of interesting, because I got to thinking about all the dumb things I’ve said....” As a current doctoral student, Derrick could conceivably be affected by things he has said using his real name on Twitter appearing in the LoC archive within his lifetime, long after he is active on the service – even after Twitter ceases to exist.

### *Some*

Most participants stated that only some tweets should be saved. Many said that the information in their tweets could be found elsewhere:

**Claire:** I don’t know that a Twitter archive would be all that useful for a coherent body of knowledge about any particular subject that you couldn’t get through other forms of searching online, in newspapers, or things like that.

Several scholars said that they used Twitter to share information they came across in their daily activities as academics; as such, they reported that their tweets often contained information that was stored in other locations. For instance, Tameka noted that “Twitter isn’t a bunch of things

that are bookmarks, it is a bunch of words around things which, if I found interesting at the time, I've already bookmarked.”

Several participants suggested that only influential people, such as politicians, celebrities, and heads of large organizations, should have their tweets archived. As Frank said, “For individual people, I don't think [we should archive] all of them.... I think that certainly important people's tweets should be, leading politicians and whatnot, should be archived.”

Many interviewees expressed concern with what we dubbed the “cheese sandwich problem” – people talking about their lunch while at their desk at work, or talking about food in general. As Sofia said, “I wouldn't necessarily save tweets about what people are eating or recipes they are thinking about using.”

Others expressed unease at having all their tweets collected for an archive without their knowledge. Twitter's Terms of Service state that “You understand that through your use of the Services you consent to the collection and use of this information, including the transfer of this information to the United States and/or other countries for storage, processing and use....”<sup>18</sup> However, several participants did not believe that they had given Twitter permission to archive their tweets:

**Dorothy:** It's in the public domain, but... I wonder about the ethics [of] harvesting the tweets without asking permission. I think maybe it should be something you sign up for, it should be an option, like you can say, ‘This is fine, my tweets [can go in] the Library of Congress archive.’”

When asked if the Terms of Service might be an appropriate place for granting permission for Twitter to allow the archiving of content posted using the service, Dorothy said, “I think it should be very, very clear and obvious, and not in small print.... Maybe you could have an option for certain tweets, just tiny little icons for [archiving].”

*All*

According to several participants, the small size of tweets and the possible value of a complete archive were worth the small cost of storage. We noticed that participants interviewed after the LoC announcement seemed more enthusiastic about the possibilities presented by a comprehensive archive. As Julio pointed out, “So much of Twitter is an interconnecting network. If you only collect some tweets and those are linking to other tweets... you may end up at a dead end if you can't access [those other tweets].” In fact, as seen in Table 1, nearly 60% of the tweets in our sample mentioned other users, illustrating that conversations on Twitter are an essential part of its use.

*Table 1: Quantitative results*

|                            | <b>Tweets contained</b>              | <b>Result</b> |
|----------------------------|--------------------------------------|---------------|
| Mention of another user    | @ <username> (anywhere)              | 58%           |
| Replies to users directly  | @ <username> (at beginning of tweet) | 36%           |
| Tweets containing links    | http://                              | 32%           |
| Link shortened with bit.ly | http://bit.ly                        | 42%           |

<sup>18</sup> “Twitter Terms of Service,” <http://twitter.com/tos>.

|                             |                                                     |     |
|-----------------------------|-----------------------------------------------------|-----|
| Link shortened with is.gd   | <a href="http://is.gd">http://is.gd</a>             | 9%  |
| Link shortened with tinyURL | <a href="http://tinyurl.com">http://tinyurl.com</a> | 7%  |
| Link shortened with ow.ly   | <a href="http://ow.ly">http://ow.ly</a>             | 2%  |
| Hashtags                    | #<word>                                             | 15% |

Some interviewees also argued that saving all tweets was the safest way to ensure that a Twitter archive would be useful in the future; as Armando said, “My bias is towards saving everything, because you never really know what people are going to need in the future. [Everything] should be archived, if it is reasonably possible.” This argument was echoed by several participants, including Frank, who added a caution: the “...difficulty of future Ph.D. students 200 years from now, sifting through, trying to find the diamonds in the haystack.”

Some people expressed that it was important to preserve the *context* of tweets in addition to their *content*. Greg explained that “A lot of the value is the context in which different things were said, and who they were said to.” Although the context of a tweet may, in fact, include the text of related tweets, it could also be comprised of information about followers, followees, and external events.

***R2. If tweets are to be archived, who do content creators think should be responsible for gathering and maintaining the collection of stored tweets?***

Opinion on who should be responsible for archiving tweets was mixed. Some scholars stated that government agencies should archive tweets:

**Derrick:** I think a governmental, or at least a nonprofit institution, would be great, in terms of having the longevity to keep this information stored. That would be a good direction to take, rather than just a private, for-profit kind of thing.

Participants indicated that government archives were an appropriate location for Twitter archives because they are generally, in Elaine’s words, “not as influenced by politics as perhaps other branches [of government].” Archives maintained by respective government agencies may have an interesting side effect; Ronnie intriguingly suggested that the LoC’s decision might stamp Twitter with additional legitimacy: “In some ways it is kind of nice because it elevates, or gives me even more reason or justification, for [using Twitter].” Dorothy agreed, noting that the LoC archive “would probably make me a more enthusiastic Twitterer, I guess out of vanity – I like that it is being archived.”

Others said that the creators should be solely responsible for archiving their own tweets, empowering the end-user. As Mitchell explained, “I think it’s very important that individuals have the opportunity to save their own content, to their own environments. This should be made as easy as possible.” Tim also suggested that content creators should be responsible, noting that the cost of saving tweets should be the burden of the content creator.

Several participants archive their own tweets, using a variety of third-party applications and other strategies to do so, such as NetNewsWire, a desktop RSS aggregator, and FriendFeed, a service to share content with others.<sup>19</sup> As Tyrone said:

<sup>19</sup> NetNewsWire, <http://netnewswireapp.com/>; FriendFeed, <http://www.friendfeed.com>.

**Tyrone:** Twitter is a private company. Someday they may just decide to just erase all tweets. I am very concerned about that, [which] is why sometimes I try to use a third party [to save] spreadsheets with all my tweets.

Several other scholars also indicated that the longevity of Twitter was questionable at best, and that the possibility of trusting the company to provide long-term access to an archive of tweets was unwise.

Most scholars objected in strong terms when asked about their academic institutions archiving their tweets. Many cited issues surrounding academic freedom; as Armando said: “My tweeting is done by me professionally, but it is not done by me as a representative of my institution.” Although the content is published on the open web, multiple scholars likened their use of Twitter to a telephone conversation:

**Ben:** It’s like my phone. If you have a phone in your office, you call people on it on a regular basis. But I would be reluctant for the university to have a policy that they would record all my phone conversations and keep them stored.

Participants also indicated that they would probably change their habits on Twitter if their institution were to archive their tweets:

**Frank:** I’d rather they did not feel obligated to archive my tweets.... I don’t feel that I am necessarily representing [my university] at all times. I certainly don’t do anything that would be controversial or cause problems for them, but I could see if I was [marked as] representing a certain university, then my tweets would be quite different. I wouldn’t want that.

Another reason that participants were wary to have their institution save their tweets was cost-related. People said that academic institutions should be focused on other priorities, “like figuring out how to deliver education without going totally broke,” as Tameka put it. In fact, several participants said that cost was the most important factor when considering responsibility for the Twitter archive; as Julio said, “There’s always a cost – who’s willing to bear that cost?”

## Findings

### *Articulating value*

The LoC plans on saving all public tweets made since 2006, when Twitter went live, and has thus chosen a model known as passive appraisal.<sup>20</sup> Saving everything that is possible to save is enticing in a system like Twitter, where the potential payoff in having such a large body of aggregate knowledge is quite great. Computational methods of extracting data from tweets have already been successful – for example, a recent study showed that Twitter activity can predict movie revenues.<sup>21</sup> Such a corpus could allow future historians to participate in the “big data”

---

<sup>20</sup> Jared A. Lyle, “Sampling the Umich. edu domain,” in *4th International Web Archiving Workshop, Bath (UK), September*, vol. 16, 2004.

<sup>21</sup> Sitaram Asur and Bernardo A Huberman, “Predicting the Future With Social Media,” *Arxiv preprint arXiv:1003.5699* (2010).



revolution; as Dan Cohen said, “Twitter is tens of millions of active users. There is no archive with tens of millions of diaries.”<sup>22</sup>

Twitter could be a perfect use-case for an all-inclusive model of appraisal, due to the low cost and negligible space to save each tweet (the corpus of public tweets is currently only about 5 terrabytes in size);<sup>23</sup> the high quality and structure of attached metadata, which saves time and effort in description; and the Twitter Terms of Service, which negate the legal need for donor agreements between the archives and content creators. However, more research and information is necessary to determine how the LoC will ingest tweets, and whether their ingest process is actually comprehensive. If not, this approach could have serious limitations.<sup>24</sup> Furthermore, our interviews indicate that some content creators do not see the value in an all-inclusive Twitter archive, preferring instead to have their tweets selectively preserved. This tension between creators and archivists is not new.<sup>25</sup> In order to respect the wishes of the creators, who are also stakeholders, and to comply with professional ethics, archivists should consider the following approaches with the Twitter archive.<sup>26</sup>

First, archivists must articulate the value of saving all tweets. Since, according to our interviews, many creators believe that the majority of tweets are trivial, noisy, and represent a deluge of data, their value as a whole should be addressed in order to appease the creators as stakeholders. Second, archivists must also consider if they have ethical obligations to the creators of content beyond what the Twitter Terms of Service state, especially since users rarely read these agreements.<sup>27</sup> Third, archivists should consider creating special collections of tweets in order to illustrate their value. For instance, collections could highlight tweets from notable celebrities, political figures, or even exemplars of “everyday folks.” Other collections could be thematic, focusing on tweets about people’s cultural foodways, election campaigns, or of academic conferences.

### **Context**

There are many types of context that contribute to the meaning of a tweet: hashtags (ad hoc tags used to identify a topic of conversation), embedded links, the identity of a tweeter, her location, and her relationship with the people with whom she is communicating on Twitter. Hashtags, links, and usernames are all contained within a tweet’s content and native metadata.<sup>28</sup> However, the events to which hashtags refer, the websites to which links point, and the real-life identity represented by a username – this information would be lost if a Twitter archive did not include extra context.

---

<sup>22</sup> Stross, “A Sea of History.”

<sup>23</sup> Campbell and Dulabahn, “Digital preservation: The Twitter archives and NDIIPP.”

<sup>24</sup> For examples of possible biases introduced by passive appraisal, see Mike Thelwall and Liwen Vaughan, “A Fair History of the Web? Examining Country Balance in the Internet Archive,” *Library and Information Science Research* 26, no. 2 (2004): 162–176.

<sup>25</sup> Sara S. Hodson, “In secret kept, in silence sealed: Privacy in the papers of authors and celebrities,” *American Archivist* 67, no. 2 (2004): 194–211.

<sup>26</sup> Society of American Archivists Council, “Code of Ethics for Archivists.”

<sup>27</sup> Nathaniel Good et al., “Stopping spyware at the gate: a user study of privacy, notice and spyware,” in *Proceedings of the 2005 symposium on Usable privacy and security*, 2005, 43–52.

<sup>28</sup> For a detailed graphic of the metadata associated with each tweet, see Raffi Krikorian, “Map of a Twitter status object,” *MeHack*, April 18, 2010, <http://mehack.com/map-of-a-twitter-status-object>.

The LoC has announced that they will only be saving tweets in their archive, and they will not systematically crawl and preserve webpages linked to from tweets.<sup>29</sup> Furthermore, since tweets can only contain 140 characters, hyperlinks within tweets are commonly truncated using services such as bit.ly, is.gd, tinyURL, and ow.ly, all of which obfuscate the location to which the link points.<sup>30</sup> The LoC has not announced whether they plan to decode these shortened links. They have also not announced whether they will make use of the Internet Archive's current project to archive these shortened URLs, called 301Works.org.<sup>31</sup>

Our quantitative results illustrate that 32% of the tweets in our sample contained links to other resources on the web (Table 1). Of these links, 60% had replaced the original URL with a truncated link from a URL shortening service. If these services stop operation, there may be no way to determine the target of the original link; over half of the links in our sample could conceivably be unviable within the LoC's Twitter archive, a prospect which warrants further attention.

Hashtags represent a similar problem. Of our sample, 15% contained hashtags (Table 1). Hashtags are often used at academic conferences so that attendees can network with one another; according to our participants, they are also used by people to follow the discussion at conferences that they were unable to attend. Unfortunately, many hashtags are difficult to understand out of context. A directory of all known hashtags on Twitter and their meanings could be extremely useful from an archival standpoint and is an area which merits further research.<sup>32</sup>

## Conclusion

This study examined the attitudes of scholars on Twitter towards archiving tweets. We found that the opinions of our 28 interviewees varied, with most people advocating that some tweets be saved for future use. People had a variety of reasons for wanting only some content preserved, including concerns about privacy and context. Participants also said that many tweets were so trivial as to be useless in the future. Halfway through our study, the LoC announced that they would be archiving Twitter; we noticed a marked difference in responses after this announcement, with many more people advocating that all tweets be saved. However, due to our data gathering and analysis methods, we are unable to provide counts or statistical information about these responses. We also found that scholars were overwhelmingly against having their tweets archived by their own university; their desires for both privacy and autonomy were cited as key reasons for this reaction.

### *Limitations and Future Research*

We used full-text search to count @replies, links, and hashtags in tweets; although this method is common in Twitter research, it slightly inflates counts.<sup>33</sup> For instance, the tweet "I'll be home

---

<sup>29</sup> Campbell and Dulabahn, "Digital preservation: The Twitter archives and NDIIPP."

<sup>30</sup> Bit.ly, <http://bit.ly>; is.gd, <http://is.gd>; TinyURL, <http://tinyurl.com/>; ow.ly, <http://ow.ly/url/shorten-url>.

<sup>31</sup> 301Works.org, <http://www.archive.org/details/301works>.

<sup>32</sup> Such as #tagdef, <http://tagdef.com>.

<sup>33</sup> danah boyd, Scott Golder, and Gilad Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *HICSS* (IEEE Computer Society, 2010), 1-10; Courtenay Honeycutt and Susan C. Herring, "Beyond Microblogging: Conversation and Collaboration via Twitter," in , vol. 0 (Los

@five” would be incorrectly interpreted as mentioning a user. Another limitation of this study is the snowball sampling method, which, while effective in allowing access to our population and acceptable for an exploratory study, hinders the generalizability of our results. A third limitation, common to all qualitative research, is the emphasis on individual experiences over broad numerical data. Future quantitative research, such as surveying, should be undertaken to broaden understanding of this area. Finally, our work used a very narrow sample; future work should expand similar investigations to other types of users and sample Twitter users more broadly in order to build more robust, generalizable results.

## References

- Asur, Sitaram, and Bernardo A Huberman. “Predicting the Future With Social Media.” *Arxiv preprint arXiv:1003.5699* (2010).
- boyd, danah, Scott Golder, and Gilad Lotan. “Tweet, tweet, retweet: Conversational aspects of retweeting on twitter.” In *HICSS*, 1-10. IEEE Computer Society, 2010.
- Campbell, Laura E., and Beth Dulabahn. “Digital preservation: The Twitter archives and NDIIPP.” Vienna, Austria, 2010.  
<http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/campbell-27.pdf>.
- Cox, Richard J., and Helen Willa Samuels. “The archivist's first responsibility: A research agenda to improve the identification and retention of records of enduring value.” *The American Archivist* (1988): 28–42.
- Crawford, Michael J. “Copyright, Unpublished Manuscript Records, and the Archivist.” *American Archivist* (1983): 135–147.
- Dunlap, Joanna C., and Patrick R. Lowenthal. “Tweeting the night away: Using Twitter to enhance social presence.” *Journal of Information Systems Education* 20, no. 2 (2009).
- Eisner, Elliot W. *The enlightened eye: Qualitative inquiry and the enhancement of educational practice*. Macmillan Pub. Co., New York, NY, 1991.
- Golbeck, Jennifer, Justin M. Grimes, and Anthony Rogers. “Twitter use by the U.S. Congress.” *Journal of the American Society for Information Science and Technology* 61, no. 8 (August 2010): 1612-1621.
- Good, Nathaniel, Rachna Dhamija, Jens Grossklags, David Thaw, Steven Aronowitz, Deirdre Mulligan, and Joseph Konstan. “Stopping spyware at the gate: a user study of privacy, Alamos, CA, USA: IEEE Computer Society, 2009), 1-10; Akshay Java et al., “Why we twitter: understanding microblogging usage and communities,” in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis* (San Jose, California: ACM, 2007), 56-65, <http://portal.acm.org/citation.cfm?id=1348556>; Edward Mischaud, “Twitter: Expressions of the Whole Self” (London, UK: London School of Economics, 2007).

- notice and spyware.” In *Proceedings of the 2005 symposium on Usable privacy and security*, 43–52, 2005.
- Heverin, Thomas, and Lisl Zach. “Twitter for City Police Department Information Sharing.” In *Proceedings of the 73rd ASIS&T Annual Meeting*. Pittsburgh, PA, 2010.
- Hodson, Sara S. “In secret kept, in silence sealed: Privacy in the papers of authors and celebrities.” *American Archivist* 67, no. 2 (2004): 194–211.
- Honeycutt, Courtenay, and Susan C. Herring. “Beyond Microblogging: Conversation and Collaboration via Twitter.” In , 0:1-10. Los Alamitos, CA, USA: IEEE Computer Society, 2009.
- Java, Akshay, Xiaodan Song, Tim Finin, and Belle Tseng. “Why we twitter: understanding microblogging usage and communities.” In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, 56-65. San Jose, California: ACM, 2007. <http://portal.acm.org/citation.cfm?id=1348556>.
- Krikorian, Raffi. “Map of a Twitter status object.” *MeHack*, April 18, 2010. <http://mehack.com/map-of-a-twitter-status-object>.
- Kvale, Steinar, and Svend Brinkmann. *InterViews: Learning the Craft of Qualitative Research Interviewing*. 2nd ed. Sage Publications, Inc, 2009.
- Letierce, Julie, Alexandre Passant, Stefan Decker, and John Breslin. “Understanding how Twitter is used to spread scientific messages.” Raleigh, North Carolina, USA, 2010.
- Lyle, Jared A. “Sampling the Umich. edu domain.” In *4th International Web Archiving Workshop, Bath (UK), September*. Vol. 16, 2004.
- McNamee, Dale. “Comment on How Tweet It Is!: Library Acquires Entire Twitter Archive.” *Library of Congress Blog*, April 19, 2010. <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/#comment-35401>.
- Mischaud, Edward. “Twitter: Expressions of the Whole Self.” London, UK: London School of Economics, 2007.
- Priem, Jason, and Kaitlin Costello. “How and why scholars cite on Twitter.” In *Proceedings of the 73rd ASIS&T Annual Meeting*. Pittsburgh, PA, 2010.
- Raymond, Matt. “How tweet it is!: Library of Congress acquires entire Twitter archive.” *Library of Congress Blog*, April 14, 2010. <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/>.
- Sanford, Matt. “Growing Around the World.” *Twitter blog*, April 8, 2010. <http://blog.twitter.com/2010/04/growing-around-world.html>.

- Schamber, Linda. "Time-line interviews and inductive content analysis: their effectiveness for exploring cognitive behaviors." *Journal of the American Society for Information Science* 51, no. 8 (2000): 734–744.
- Society of American Archivists Council. "Code of Ethics for Archivists," February 5, 2005. [http://www.archivists.org/governance/handbook/app\\_ethics.asp](http://www.archivists.org/governance/handbook/app_ethics.asp).
- Stross, Randall. "A Sea of History: Twitter at the Library of Congress." *The New York Times*, May 1, 2010, sec. Business. [http://www.nytimes.com/2010/05/02/business/02digi.htm?\\_r=1](http://www.nytimes.com/2010/05/02/business/02digi.htm?_r=1).
- Stutzman, Fred. "Twitter and the Library of Congress," April 14, 2010. <http://fstutzman.com/2010/04/14/twitter-and-the-library-of-congress/>.
- Thelwall, Mike, and Liwen Vaughan. "A Fair History of the Web? Examining Country Balance in the Internet Archive." *Library and Information Science Research* 26, no. 2 (2004): 162–176.
- "Twitter Terms of Service." <http://twitter.com/tos>.
- Williams, Evan. "The Evolving Ecosystem." *Twitter blog*, September 2, 2010. <http://blog.twitter.com/2010/09/evolving-ecosystem.html>.
- Young, Jeffrey R. "10 High Fliers on Twitter." *The Chronicle of Higher Education*, April 10, 2009.
- Zeldman, Jeffrey. "You cannot copyright a tweet, particularly the comments section." *The Daily Report*, February 25, 2010. <http://www.zeldman.com/2010/02/25/you-cannot-copyright-a-tweet/>.
- Zhang, Yan, and Barbara M. Wildemuth. "Qualitative analysis of content." In *Applications of social research methods to questions in information and library science*, 308-319. Westport, Connecticut: Libraries Unlimited, 2009.
- Zhao, Dejin, and Mary Beth Rosson. "How and why people Twitter: the role that micro-blogging plays in informal communication at work." In *Proceedings of the ACM 2009 international conference on Supporting group work*, 243-252. Sanibel Island, Florida, USA: ACM, 2009. <http://portal.acm.org/citation.cfm?id=1531710#CIT>.

## Appendix A: Recruitment Web page

### Academics on Twitter Needed for Research Study

*Are you a faculty member, post-doc, or doctoral student? Do you use Twitter at least once a week? Would you like to participate in a research study about academics on Twitter?*

This research study investigates the reasons that academics use Twitter. This research study investigates your thoughts about academic impact and long-term value on Twitter. If you choose to participate in this research study, you will be interviewed by a researcher over the telephone at your convenience. The whole research study will take between 30 and 45 minutes of your time.

Oversight for this study is provided by the UNC Chapel Hill Institutional Review Board for Social and Behavioral Research. If you have questions or concerns about this study please contact the IRB at 919-966-3113 or by email at [irb\\_questions@unc.edu](mailto:irb_questions@unc.edu). You can also contact the faculty advisor, Barbara Wildemuth, at [wildemuth@ils.unc.edu](mailto:wildemuth@ils.unc.edu), with any questions or concerns about this study.

For more information or to take part in this research study please contact

Kaitlin Costello by e-mail at [kaitcost@email.unc.edu](mailto:kaitcost@email.unc.edu)  
or by direct message on Twitter to [@k8lin](https://twitter.com/k8lin)

or

Jason Priem by e-mail at [priem@email.unc.edu](mailto:priem@email.unc.edu)  
Or by direct message to Twitter to [@jasonpriem](https://twitter.com/jasonpriem)

Participants should use Twitter at least once a week and hold an academic appointment (faculty members, including adjunct or emeritus; post-doctoral research appointments; or doctoral students) at a university or college.

## Appendix B: Semi-structured interview guide

*The numbered questions will be posed to the subjects. Indented points are examples of follow-up questions that may be asked, if appropriate and relevant to the discussion generated by the main questions. Additional follow-up questions may be asked that are not present in the guide, depending on the responses from subjects.*

1. What is your role or position at your University (faculty, post-doc, PhD student)?
  - a. How long have you held this position?
  - b. What academic discipline are you in?
2. Why do you use Twitter?
3. Can you tell me about how you use Twitter?
  - a. How often do you post on Twitter?
  - b. How many people do you follow?
  - c. What kinds of people do you follow, and what do they tweet about?
4. Do you post titles or links to scholarly articles on Twitter?
  - a. Do the people you follow post links to or titles of scholarly articles?
    - i. How often do you read the scholarly articles people tweet about?
5. *If the answer to question 4 is yes...* Compare referencing an article on Twitter to citing that article in your scholarly work.
  - a. Do you tweet about academic articles that you wouldn't cite in your academic work?
  - b. What are some similarities between referencing an article on Twitter and citing that article in your scholarly work?
  - c. What are some differences between referencing an article on Twitter and citing that article in your scholarly work?
6. What impact do the tweets from other academics in your Twitter feed have on your academic work?
  - a. What about your own tweets? What impact do they have on your own academic work?
7. **Do you think that some or all of your tweets have lasting value?**
  - a. **What about the tweets of people you follow?**
8. **Do you think that some or all of your tweets should be saved so that they can be viewed in the future?**
  - a. **For how long do you think your tweets should be accessible?**
  - b. **Which tweets of your own do you think should be saved?**
9. **Do you think that some or all of the tweets that you follow should be saved for future use?**
  - a. **Which tweets of others do you think should be saved?**
10. *If the answers to questions 7-9 are yes...* **Who do you think should be responsible for making sure that tweets are made available in the future?**
  - a. **Should your institution play a role in this activity?**
11. **Did you have any other comments to share?**
12. *If we have not yet obtained enough participants...* **Would you mind tweeting about this interview to help us recruit more participants?**
  - a. *If yes, offer to e-mail them the text for the tweet*

Thank you very much for your time. We really appreciate your participation.