Attitudes about Institutional Archiving of Social Media

Catherine C. Marshall; Microsoft Research Silicon Valley; Mountain View, CA 94043
Frank Shipman; Department of Computer Science, Texas A&M University; College Station, TX 77843

Abstract

This paper compares the results of two surveys that characterize attitudes about the institutional archives of two forms of public social media, Twitter tweets (i.e. microblog posts) and Flickr photos (i.e. shared visual media). Internet-savvy respondents were asked to assess three statements about a hypothetical scenario in which the Library of Congress archived and provided access to the social media in question. Access to the archives varied in three ways: (1) the public was given immediate access to the archive; (2) researchers were given immediate access to the archive; and (3) the public was given deferred access (by 50 years) to the archive. We found that access to photos was received best when it was deferred by 50 years, and access to tweets was received best when it was limited to researchers, hence suggesting that institutions pay careful attention to access limitations when they are seeking public acceptance of their archiving efforts.

Introduction

Social media, by its very nature, introduces questions about content ownership. Digital material is uploaded by individuals; it may document social or group events; it is shared with and commented on by an extended network of friends, family, and associates; and it is curated in practice through a combination of benign neglect and the uneven application of personal information management techniques. Yet increasingly people see services like Facebook as the first place to go to deposit and share socially meaningful personal artifacts, from everyday accounts of their lives to wedding pictures and online memorials. In fact, according to an IDC annual survey published in May, 2010, 70% of the world's digital content is now created by individuals [1].

It is no wonder then that there is little resolution (beyond the obligatory End User License Agreements that users seldom read) about who owns—and, indeed, who is responsible for—the social media that is a vast and growing component of the Internet. End User License Agreements (EULAs) may even sidestep the matter. For example, Facebook's EULA asserts:

> "You grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post on or in connection with Facebook… This IP License ends when you delete your IP content or your account *unless your content has been shared with others, and they have not deleted it*." (italics added by the authors)

In other words, Facebook cannot really circumscribe the conditions under which their license holds, since users' content may migrate into the purview of their friends.

Content ownership comes into play most crucially when we design services and applications to archive, reuse, remix, or remove social media. We have been investigating social media ownership issues using a series of Mechanical Turk surveys that probe respondents' current attitudes and practices; the surveys combine open-ended questions about use with realistic scenarios that test respondents' attitudes in specific situations [2, 3].

One scenario we have included in our series of surveys poses questions about institutional archiving, since people may have difficulty meeting personal archiving challenges as individuals, and there may be social value (for example for historians, anthropologists, and other types of social scientists) in creating a more broadly accessible store of aggregated personal content.

Specifically, we were interested in learning how respondents view real and hypothetical efforts by the Library of Congress to collect, store, and potentially provide access to personal media. Are people more cautious (or less cautious) about granting memory institutions the authority to archive their digital assets—for example, Twitter tweets, digital photos and videos, game-related materials and machinima, product reviews, or audio recordings—than they are individuals? If institutional archiving is implemented, who can see and use personal digital assets? Does time factor into more universal access, or should access be limited to historians and researchers?

In this paper, we describe two studies we have performed to probe respondents' reactions to institutional archiving of public Twitter microblogging content and public Flickr photos. We briefly discuss the method used to conduct the surveys and summarize the respondents' demographic characteristics and how they compare to the larger population. Finally, we present and compare the results of the two surveys' probes into this hypothetical institutional archiving effort.

Method

To perform this study, we sought respondents who were familiar with social media in general, and who were Internet-savvy without being part of the industry; the study population needed to be familiar with the particular type of social media in question. In the first survey, we screened for familiarity with (and participation in) Twitter, the microblogging service, and in the second survey, we screened for use of Flickr, or some equivalent digital photo sharing site.

Mechanical Turk. We used documented best practices to implement the survey as a Mechanical Turk HIT (Human Intelligence Task). Mechanical Turk (MT) is an online labor marketplace wherein qualified workers accept online tasks; the service tracks worker identity (while concealing it, ensuring anonymity) and reliability. The general demographics of the MT community, including those of US-based workers, have been published elsewhere [4].

Mechanical Turk thus gave us access to a diverse, yet reliable, population of English-speaking respondents who had the desired demographic profile. Researchers have identified ways to test the validity of responses (e.g. through the use of a circumscribed number of reading comprehension questions [5]); we also made certain the respondents had spent a realistic amount of time on survey to answer the questions. Our methods for detecting fraudulent participation and for ensuring response quality are described in [2, 3]. The quality of responses to open-ended questions reassured us that the respondents took the questionnaires seriously and gave us confidence in the data we gathered this way.

Survey structure. The two surveys we administered were structured in three parts. First we gathered basic demographic information, including two open-ended questions in which the participants told us what kind of activities they engaged in on the Internet, and what kind of materials they published and shared online. This information was useful in characterizing the respondents and comparing them with larger populations.

The second part of both surveys presented the respondents with several realistic scenarios, and requested that they rate some statements about the scenarios on a 7-point Likert scale. These statements would enable us to assess their attitudes. In this paper, we focus on two equivalent hypothetical situations and the three associated statements that were presented in both surveys.

Finally, the surveys (particularly the second survey about photo-sharing) included some questions about the respondents' own practices. Although the first survey only included a modest number of these, respondents seemed more willing than we expected to answer this type of question, and to answer the questions expansively and—by some measure—authentically and candidly. Thus the second survey included many more questions of this sort. For the sake of brevity, we do not discuss this type of result here, but rather refer the interested reader to our other papers.

Both surveys were deployed for two weeks. The microblogging survey garnered 190 responses, 173 of which passed our reliability tests, and the photo-sharing survey garnered 250 responses, 242 of which passed our reliability tests.

Respondent characteristics. In both surveys, the participants were generally in their 20s and 30s (as we would expect, given our screening criteria, and the demographic profile of the MT service), although they ranged much more broadly. Like most surveys (and the service itself), females are slightly overrepresented in the microblogging survey (at 61%) and more so in the photo-sharing survey (at 72%). In both surveys, students made up about 1/3 of the respondents. The majority survey-takers had finished college, a characteristic which accurately reflects the composition of English-speaking Turkers. More than ½ of the survey population have more than 10 years of experience using the Internet, and almost all of the respondents have more than 7 years of Internet experience.

Thus our participants are generally young (in their 20s and 30s), college educated, Internet savvy, but as we learned from the open-ended questions, represent a fairly diverse swath of the online population.

The open-ended questions revealed an extensive range of types of social and individual Internet use. Individual uses included reading (e.g. news, blogs, websites); research (reflecting the respondents' own interests and behalf of others); working (in addition to being Turkers, many respondents worked on independent projects and intellectual piece-work); shopping (including specific mention of sites such as Amazon, eBay, and the crafting site etsy); and consuming media (e.g. watch YouTube, TV, or movies). Social uses included participation in online communities (e.g. DeviantArt, TinierMe); social networking (especially Facebook); online gaming (including massively multiplayer online roleplaying games such as the World of Warcraft); and the use of communication tools such as Skype or chat programs. In summary, although the respondents are savvy, they are not dominated by a single interest or type of use. We were very much interested in this sort of diversity of purpose and interests.

Privacy concerns. Our general open-ended questions about publishing and sharing (which did not include any direct questions about privacy) revealed that respondents had a variety of privacy concerns that arose organically. The responses ranged from very privacy-aware to cavalier. In other words, some respondents cited privacy as a central concern in what they decide to share or what privacy controls they use; others did not bring up privacy at all; while still others thought of it as damage-control, or as sharing as something they do in spite of their own best interests. As we would expect from the literature [6], respondents self-report deviations between their attitudes and what they actually do (other contradictions just appear in respondents' answers). For example, a respondent in the photo-sharing survey said, "Some of the information that I put on the Internet is private. I tend to put more than is probably safe."

Privacy-aware practices reveal that some respondents think of certain types of content as being more private than others, but that there is no consensus about which types of content are private. For example, contrast the following two responses:

> [1] *"...I keep profile pictures, but I don't keep a bunch because of worries about my privacy. I worry of them archiving photos, so a lot of photos I might share I don't.*"
>
> [2] *"...On more public forums I am very general, I do not use my name or give any locations.*"

Respondents also reported being aware of privacy concerns, but deliberately defying them (e.g. "*I share everything except my home address. I dont have anything to hide so Im not afraid of anyone*"). They also report changes in practice that acknowledge a growing awareness of the perils of sharing (e.g. "*I used to share all of my information online and hold nothing back but more recently I have limited this sharing to just my name, location, and pictures of myself and my family*").

Because these distinctions seemed important, we independently coded all respondents into three categories: privacy-aware, social-sharer, or neutral, based on their responses to the open-ended questions. We then compared our independent coding efforts, and only used the results when we agreed. By these conventions, 26 respondents were privacy-aware and 62 were social-sharers. These privacy concerns are discussed in the results we report later in this paper.

Relevant Scenarios

Although the questionnaires contained multiple related scenarios and statements about the scenarios, in this paper we focus on two scenarios, which elicited comparable responses. They were parallel institutional archiving scenarios that covered the media type in question (Twitter tweets or the public digital photos in Flickr, which at the time of the survey contained around five billion images). The scenarios were cribbed from the Library of Congress's acquisition of the Twitter archive last summer, which was greeted with a surprising amount of acrimony in a vocal community of Twitter users, and tested the circumstances under which these archives could be used. In other words, if a large public institution acquired a major social media resource, would people be comfortable if access were given to the general public (note that we were clear to specify that this material is the currently public portion of the resource).

In addition to asking this general question, we modified the hypothetical access in two ways that were parallel in both scenarios. First we limited the access to the tweets or the pictures to researchers (rather than to the general public). Then we specified that access would not be limited, but would be deferred for 50 years.

Table 1 summarizes the scenarios, the test statements, and the mean values for the six possible situations we tested. The distribution of responses to many of the statements is bimodal, thus we are only including the means in this table to give the reader a preliminary sense of the relative strength and favorability of the responses. The next section gives a more thoughtful breakdown of the results.

Table 1. Summary of two parallel institutional archiving scenarios, related statements, and their mean responses

Statement	**Twitter (mean)**	**Flickr (mean)**
Scenario: The Library of Congress is acquiring the public portion of the resource (Twitter/Flickr)		
The Library of Congress can give everyone access to the archive.	4.22	4.14
The Library of Congress can give researchers access to the archive.	4.72	4.65
The Library of Congress can give everyone access to the archive after 50 years has passed.	4.59	4.93

DISCUSSION OF RESULTS

Today, many of our personal digital assets are stored in social media services. The Pew Research Center estimates that 2/3 of Americans store personal data in the cloud [7], and that almost half of Americans are social networkers [8]. Furthermore, the digital materials that individuals store locally are in some state of disarray; many of them are maintained through a haphazard combination of intentional practices and benign neglect. Many people may be unaware of what they have, where they have put it, and why they have kept it [9]. It is little wonder that the acquisition of some of the larger shared resources such as Twitter, Flickr, and even Facebook by public institutions seems like a good idea, and perhaps the most reliable way of ensuring that some portion of the important records of our lives as we live them today remain viable from a historical perspective.

Yet at the outset of these major undertakings, public reaction seems mixed. Are these shared resources perceived as ephemeral? Are they guaranteed some measure of privacy through obscurity? Are they too personal for a public institution to control? Are there some policies and measures that may be taken to mitigate against the most controversial of these actions? We can examine our data to construct an initial snapshot of public attitudes and how they compare across different access limitations and media types.

First, it is important to look at the most general scenario: The Library of Congress acquires the public portions of these resources (either the entirety of public tweets from Twitter, or the entirety of the public photos on Flickr). Both are very large collections. Tweets seem to be regarded as more ephemeral than photos, although certainly they have been examined as windows onto important geopolitical events of the day [10]; tweets are also the target of several emerging archiving services (e.g. Twapper Keeper and BackupMyTweets, services which may currently violate Twitter's Terms of Service), as well as other services that are indicative of a less-than ephemeral take on the content (e.g. Twournal, which allows a user to automatically generate a book from his or her tweets). Flickr, on the other hand, is regarded as a significant store of digital images; however, many Flickr users regard even their public photos as personal and therefore possibly a sensitive target for a public archive [11].

Figure 1 shows a comparison of reactions to the statement, "The Library of Congress can give everyone access to the [Twitter/Flickr] archive." Figure 1 compares the percentage of respondents in the two surveys who have assigned each of the seven possible values to the statement (where 1 is "disagree strongly" and 7 is "agree strongly"). Note that the two graphs are similar in their basic contours; the distribution of responses is bimodal, which indicates that the statement is controversial. Some respondents would clearly welcome institutional archiving efforts and the subsequent public access this effort ensures, while others feel that this access is inappropriately broad. It is interesting to note that Twitter users have a stronger very negative response than Flickr users do; we might surmise that privacy interests are at the root of this response, since some Twitter users invoked the privacy through obscurity reasoning when they initially reacted to the real Library of Congress effort announced last summer. It is also evident that respondents feel less strongly about a comparable effort (hypothetical as far as we know) to archive and provide access to personal (but still public) photos. There was no statistically significant difference in the responses to the two scenarios (Mann Whitney, p=.92).

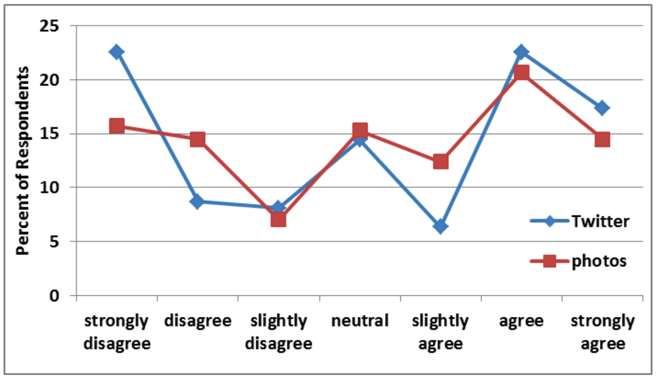

***Figure 1.** Comparing open access to public photos and tweets in the hypothetical Library of Congress archive*

The second condition we tested was to limit access to the hypothetical archive to researchers. Restricting access in this way seemed to reduce the negative reactions to such an undertaking;

the two graphs, showing in Figure 2, are very similar, perhaps because respondents don't see the reuse possibilities (or other violations of privacy) so readily cropping up. We know from the open-ended responses to the photo-sharing survey that some respondents feel very strongly that open access to photos provides limitless avenues for abuse, that the Internet is a digital Wild West (several respondents described it in exactly these terms). It is interesting that this graph shows the least difference between media types; limiting access to researchers denatures the strongest objections, and possibly dampens respondents' enthusiasm as well (since they possibly will not be taking advantage of such a store). As with the first scenario, there was no statistically significant difference in the responses (Mann Whitney, p=.99).

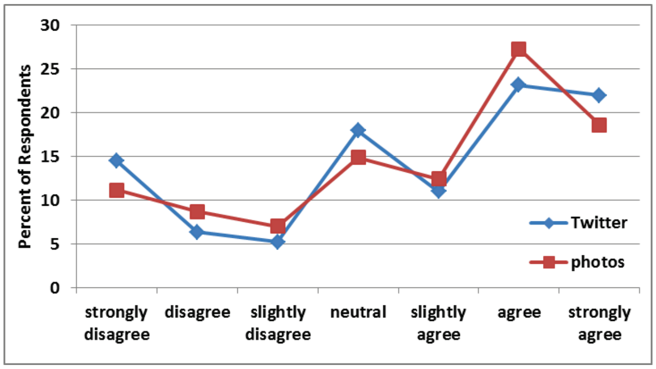

Figure 2. *Comparing access limited to researchers in the hypothetical Library of Congress archive*

What of the third variation, deferring access for 50 years (which outpaces most respondents' current life spans)? Here we see a statistically significant difference in the values respondents' assigned to the different media types (Mann Whitney, p<.02). More importantly, respondents to the Twitter survey find access in 50 years controversial; respondents to the photo-sharing survey do not. There are fewer strongly negative reactions (as well as more strongly positive reactions) to deferred access to a photo archive; this may well indicate that respondents perceive the value of such an archive (perhaps even the value to themselves of an institutional archive of personal photos), and that this value outstrips the negative aspects of privacy violation.

Given the differences evident in the reactions to these hypothetical efforts, we might surmise that the public may respond better to a variety of access limitations. It seems important to understand what underlies these differences, probably using methods that dig more deeply into individual motivations and practices.

One hypothesis that we tested with the photo study data was that demographic factors might influence the results. For example, the mean values for men trended slightly higher for both current and delayed public access to a hypothetical Library of Congress photo archive, but the difference did not turn out to be statistically significant. Similarly, younger people (born in the 1980s or later) were slightly more amenable to all three modes of access to a hypothetical Library of Congress photo archive than their older counterparts were (respondents born in the 1970s or earlier), but again, none of these differences were significant. As we would expect from the age-based results, students were also slightly more amenable to all three modes of access to a hypothetical Library of Congress photo archive, but again, we found no statistical significance.

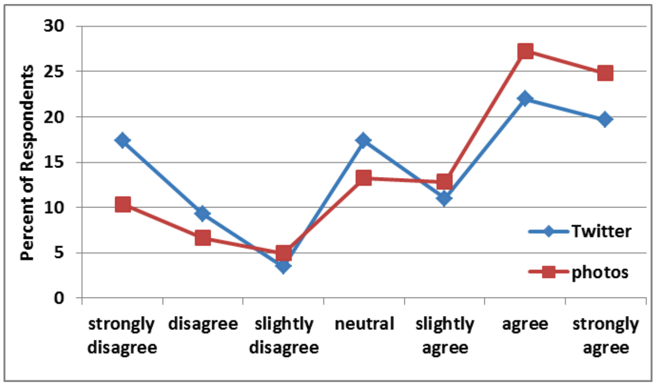

Figure 3. *Comparing access deferred for 50 years in the hypothetical Library of Congress archive*

On the other hand, we did discover that separating the privacy aware respondents from the social sharers in the Twitter study generated a near-significant difference (Mann Whitney, .1<p<.05) for the open access and researcher access Library of Congress questions, but not for delayed access. This difference is likely to be fully significant if our sample size increased (by the coding conventions discussed earlier, 26 respondents were privacy aware and 62 were social sharers). Table 2 shows these emerging trends.

Table 2. Comparison of emerging trends in hypothetical Twitter archive access modes between privacy aware and social sharer respondent categories

Privacy attitude	**Unlimited access to Twitter archive**	**Access to Twitter archive limited to researchers**	**Access to Twitter archive deferred for 50 years**
Privacy aware (26)	3.23	4.04	4.62
Social-sharer (62)	4.10	4.98	4.56

FUTURE WORK

Our findings to-date suggest that institutional archiving (in our scenarios, performed by the Library of Congress) is better accepted if access to the materials is either restricted to researchers (as we found in our Twitter survey), or if the content is off-limits to the general public for 50 years after it has been collected from a social media website (as we found to be acceptable for shared photos).

Our future work will focus on further studies of the sort we described in this paper, using scenarios and practice-driven questions to explore attitudes and behavior associated with other media types. Media types slated for future study include structured content (such as reviews), audio content (e.g. podcasts, Skype files), game content (from Massively Multiplayer Role Playing Games such as World of Warcraft), and finally, heterogeneous social networking content, including explicit representations of users' social networks and the mixed media they share. We are planning to perform additional analyses on the data we collect to identify interesting differences in the behavior and attitudes of students and non-students, and age- and gender-specific practices; we are also exploring the inclusion of other demographic variables.

Other types of qualitative studies are also planned so we can delve more deeply into practice-related questions that it is difficult to pin down using survey methods. Privacy-related questions and certain controversial reuse practices, for example, are more effectively investigated using interviews and observations.

References

[1] R. Wray, Goodbye petabytes, hello zettabytes. *UK Guardian*, 3 May 2010.

[2] C.C. Marshall and F.M. Shipman, Social Media Ownership: Using Twitter as a Window onto Current Attitudes and Beliefs. To appear *Proc. CHI'11*, Vancouver, BC, May 7-12.

[3] C.C. Marshall and F.M. Shipman, The Ownership and Reuse of Visual Media. To appear Proc. JCDL'11, Ottawa, Canada, June 13-17, 2011.

[4] P. Ipeirotis, The New Demographics of Mechanical Turk. http://behind-the-enemy-lines.blogspot.com /2010/03/new-demographics-of-mechanical-turk.html.

[5] A. Kittur, E. Chi, and B. Suh, Crowdsourcing User Studies with Mechanical Turk. *Proc. CHI'08*, pp. 453-456.

[6] A. Acquisti and J. Grossklags, Privacy Attitudes and Privacy Behavior, in J. Camp and S. Lewis (Eds.) *The Economics of Information Security*, Kluwer, Boston, pp. 165-178.

[7] J. Horrigan, Use of Cloud Computing Applications and Services, Pew Internet and American Life Project, http://www.pewinternet.org/ Reports/2008/Use-of-Cloud-Computing-Applications-and-Services.aspx, Sept. 12, 2008.

[8] L. Rainie, K. Purcell, A. Smith, The Social Side of the Internet, Pew Internet and American Life Project, http://www.pewinternet.org/ Reports /2011/The-Social-Side-of-the-Internet.aspx, Jan. 18, 2011.

[9] C.C. Marshall, S. Bly, F. Brun-Cottan, The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives. *Proc.Archiving 2006*, IS&T, Springfield, VA, 2006, pp. 25-30.

[10] A.L. Hughes and L. Palen, Twitter adoption and use in mass convergence and emergency events. *Int. Journal of Emergency Management*. 6 (3/4) 2009, pp. 248-260.

[11] A. Besmer and H.R. Lipford, Moving Beyond Untagging: Photo Privacy in a Tagged World. *Proc. CHI'10*, pp. 1563-1572.

Author Biography

Cathy Marshall is a Principal Researcher at Microsoft Research, Silicon Valley. A list of her publications can be found at http://www.csdl.tamu.edu/~marshall/pubs.html.

Frank Shipman is a Professor of Computer Science and Associate Director of the Center for the Study of Digital Libraries at Texas A&M University. His website is http://www.csdl.tamu.edu/~shipman/.