On the Institutional Archiving of Social Media

Catherine C. Marshall
Microsoft Research, Silicon Valley
1065 La Avenida
Mountain View, CA 94043
1-650-693-1308
cathymar@microsoft.com

Frank M. Shipman
Department of Computer Science
Texas A&M University
College Station, TX 77843-3112
1-979-862-3216
shipman@cs.tamu.edu

ABSTRACT

Social media records the thoughts and activities of countless cultures and subcultures around the globe. Yet institutional efforts to archive social media content remain controversial. We report on 988 responses across six surveys of social media users that included questions to explore this controversy. The quantitative and qualitative results show that the way people think about the issue depends on how personal and ephemeral they view the content to be. They use concepts such as creator privacy, content characteristics, technological capabilities, perceived legal rights, and intrinsic social good to reason about the boundaries of institutional social media archiving efforts.

Categories and Subject Descriptors

H4.3 Information Systems: Communications Applications.

General Terms

Design, Experimentation, Human Factors.

Keywords

Library of Congress, archiving, information rights, survey.

1. INTRODUCTION

Several years ago, as part of the National Digital Information Infrastructure Preservation Program (NDIIPP), the Library of Congress held workshops about information permanence with students, representatives of a first generation of so-called digital natives. When one of these students was asked what she thought should be saved, she speculated, "*They should just save Facebook. That is our generation's scrapbook, yearbook, Guinness World Record.*" [15] Indeed, social media plays an important function in how many of us negotiate our everyday lives, in how news and gossip are disseminated, and in how we record personal events large and small. It is no wonder that in 2010 the Library of Congress embarked on a project to archive public tweets after Twitter, a privately-held corporation, donated this content; arguably tweets can be an irreplaceable resource for documenting trends, breaking news, public reactions, and the more mundane aspects of peoples' lives [10].

Yet the project was not without controversy. Not only was there consternation in some quarters about the potential worthlessness of most tweets, but also there was outcry about the effect this acquisition would have on personal privacy. In his blog, social media privacy researcher Fred Stutzman posed a scenario that illustrated why we might be concerned:

> *"This is what makes Twitter's "gift" troubling. It assumes that all content shared publicly is truly public and for posterity. ... [Consider this scenario:] Bob wants to be practically obscure – private in public – without going to all the trouble of setting up complicated privacy controls. So what happens, two years from now, when Bob accidentally discloses his handle in the wrong context, and he needs to remove some Tweets?"* [24]

Erosion of personal privacy is only one aspect of the institutional archiving of social media that makes it a complex endeavor. Most transfer of personal materials to institutions used to be done by explicit design, though donor agreements that linked individuals and their families with specific archives or libraries that had the resources to maintain such collections, and possibly to provide future generations of historians and researchers access to them.

Expectations have changed. The curation of personal materials has become more difficult [21]; individuals have orders of magnitude more digital stuff than they did print media and physical belongings. They are seldom fully aware of what they have and where they've kept it. Furthermore, the ownership and control of these digital belongings is significantly more entangled: individuals may no longer own what they have stored online, nor are their assets centralized. Instead, digital belongings may be subject to the terms and conditions of many different agreements and competing legal and social interests [23].

Specifically, users of social media services such as Twitter and Facebook agree to terms and conditions that put the content they contribute into the hands of the service providers; thus the content is legally owned and controlled by these corporations as part of their assets. Because the content is shared socially and may be re-used more broadly, its de facto ownership is fluid, and its reach may extend far beyond a contributor's original intent.

Moreover, the content has potential archival value as it moves through time and space. A personal tweet from a day of rioting may become part of a larger exchange that reflects a country's growing political unrest; a vacation photo of a landmark may become a reference image that is used to illustrate dozens of minor publications; a book review, movie rating, or crowdsourced biographical information may become metadata for a resource like the Internet Movie Database (IMDB). As this kind of value grows and is recognized, a service's content may be acquired by a public institution such as the Library of Congress, a non-profit such as the Internet Archive, or a for-profit corporation that perceives the value of the service's content (for example, consider Yahoo's acquisition of Flickr, the photo-sharing service).

In spite of general recognition of the historical or cultural value of such acquisitions by public institutions, as we saw earlier, they are not without controversy. Just because a user cannot figure out Facebook's privacy controls, are her posts and profile public? To

JCDL'12, June 10–14, 2012, Washington, DC, USA.
Copyright 2012 978-1-4503-1154-0/12/06…$10.00.

whom do they belong when she stops using the service? If a friend copies a photo, does she lose control of it into perpetuity?

We performed a series of six scenario-based surveys to explore the attitudes of Internet-fluent (but not necessarily technologically trained) people to the issues that arise when public institutions acquire social media. The surveys also cover other aspects of the ownership and control of individually-contributed online content, including respondents' own practices, but in this paper, we focus on the results that pertain to institutional archiving of social media. In addition to identifying emergent issues, we sought answers to the following questions:

(1) How pervasive are the attitudes that we observed when the public Twitter feed was donated to the Library of Congress?

(2) Do different media types raise new concerns for the institutions doing the archiving?

(3) Are there ways an institution can mitigate the anxiety that may arise? (e.g., will limiting access help?)

We begin this paper by describing the surveys themselves; we especially focus on three recent surveys that were aimed at eliciting explanations of what the respondents perceive to be off-limits for institutional archiving efforts. We then summarize salient results from earlier surveys and discuss new results from the most recent surveys. Finally we explore the implications of the cumulative results for future institutional archiving efforts.

2. SIX SURVEYS: METHODS AND DATA

In this paper, we discuss and compare data collected in six different scenario-based surveys; a total of 988 valid responses (out of 1060 total) form the basis for our findings. Each survey covers a significant genre of user-contributed content, including tweets, photos, reviews, recorded videoconferences, podcasts, and educational media.

2.1 Using Mechanical Turk

The surveys were conducted using Mechanical Turk to solicit and screen respondents and to collect the data. We limited prospective respondents to the English-speaking US-based Mechanical Turk community for three reasons: First, limiting the respondent pool simplified justifiable concerns about scenario comprehension and legal and cultural norms; in other words, we wanted to ensure that the situations we presented were understood in comparable ways to provide us with a firm basis for data analysis. Second, past studies have confirmed that US-based Turkers participate in the community out of interest and for entertainment, rather than for meaningful compensation [8]. This motivation is evident in their lucid and detailed responses to open-ended questions; it is unlikely they were influenced by the small payment. Finally, US-based Turkers have demographic characteristics that suggest they are representative of a larger, fairly diverse population of Internet users who are fluent in social media; we felt Mechanical Turk was a good way to reach a broad set of respondents.

Our own previous experiences with Mechanical Turk, coupled with documented best practices [4][9][10], enabled us to screen respondents and assemble reliable high-quality datasets. In particular, we made certain that respondents performed adequately on reading comprehension questions embedded in each survey; that they spent sufficient time completing the surveys; and that they responded fully to the open-ended questions that formed an important qualitative component of the data. We paid respondents for survey completion at a level commensurate with other similar Mechanical Turk tasks; respondents were paid regardless of whether their data was discarded.

2.2 Surveys and Respondents

This paper describes the results of the institutional archiving portions of four unpublished surveys, in addition to summarizing and contrasting these new results with results from two earlier studies of microblog entries (Twitter tweets) [16] and personal photos (photos with identifiable human subjects) [17]. The methods sections of the earlier papers contain additional details about the implementation of the surveys and how we ensured data reliability. The new surveys build on and refine the methods used in the early surveys. Most significantly, we added two types of open-ended questions: we asked about respondents' *current practices* (focusing on specific instances of media creation and reuse) and about their *ethical concerns* with social media reuse and archiving. We also added three practice-based questions that we felt would be good indicators of respondents' current privacy behavior; we are aware that abstract questions about future privacy behavior are apt to yield aspirational results [1][3].

Table 1 summarizes the surveys we have fielded to-date. The table documents the number of responses we received for each survey, how long it took respondents on average to complete the survey, and some key demographic properties of each set of respondents. Since the surveys were offered over comparable durations (two weeks of active responses, until the survey reached 250 participants), and we required respondents be users of the survey's social media type, the number of responses is likely to reflect the relative popularity of the media type. Hence photos are probably the most familiar media type among the prospective respondents, and recorded videoconferences are the least.

Table 1. Summary of survey respondents and demographics

Survey Media	screened responses (out of total)	average completion time	% female/ male/ no response	have college degree	have some college	current students	born before 1960	born 1960-1969	born 1970-1979	born 1980–1989	born after 1989
Twitter	173 (of 190)	8 minutes 44 seconds	61/39/0 (105/68/0)	54% (94)	88% (152)	did not ask	4% (7)	4% (7)	17% (29)	64% (110)	11% (19)
Photos	242 (of 250)	13 minutes 7 seconds	71/27/1 (173/66/3)	55% (133)	91% (221)	34% (82)	2% (4)	10% (25)	22% (53)	56% (135)	10% (23)
Reviews	203 (of 216)	14 minutes 23 seconds	59/41/0 (119/84/0)	62% (125)	92% (186)	32% (64)	3% (7)	12% (25)	23% (47)	50% (101)	9% (19)
Podcasts	170 (of 180)	11 minutes 30 seconds	46/53/1 (78/90/2)	59% (101)	91% (154)	32% (55)	2% (4)	9% (15)	22% (37)	52% (88)	15% (25)
Videos	98 (of 107)	10 minutes 7 seconds	55/43/2 (54/42/2)	72% (71)	94% (92)	24% (24)	5% (5)	8% (8)	26% (25)	47% (46)	13% (13)
Educational recordings	102 (of 117)	11 minutes 52 seconds	57/43/0 (58/44/0)	65% (66)	96% (98)	31% (32)	8% (8)	11% (11)	21% (21)	50% (51)	9% (9)
Total	988 (of 1060)		59/40/1 (587/394/7)	60% (590)	91% (903)		4% (35)	9% (91)	21% (212)	54% (531)	11% (108)

If we look at the relative demographic participation for each survey, we can pinpoint certain distinctions: The photo survey has the greatest proportion of female respondents (at 71%) and the podcast survey has the least (at 46%). As is true of surveys in general, and Mechanical Turk surveys in particular, women tend to participate in higher numbers [8]. Current students constituted about a third of the participation in most of the surveys, with the exception of the videoconference survey. We believe this can be attributed to the fact that we limited participation to Turkers familiar with videoconferences (rather than Turkers familiar with Skype); upon reflection, the term ‘videoconference’ is associated with the workplace, which may discourage prospective respondents who use Skype (or a similar application) to chat with their friends and families. The surveys were dominated by respondents in their twenties and thirties, although they all had participants who were over fifty and under twenty.

The surveys identified strong participation in diverse types of social media creation and use. Some respondents limited their contributions to basic profile information, “*name, address, email, phone number, bank info for shopping, anything I need to basically*" [EDU106]; others focused on fairly specific topics and genres, “*Labor law articles for work*” [EDU103]. Still others seemed fairly expansive, and revealed the respondent’s interests and media creation skills: “*Music oriented information, including, music and music gear reviews, videos on how to work musical gear, and reviews on music*” [EDU257]. This broad sample of Internet users was important to us.

The six surveys each had a parallel structure. The first part of the survey elicited demographic and background information about the respondent. The second set of questions was scenario-driven. Respondents read about a social media situation. Then they were presented with a series of statements, and were asked whether they agreed or disagreed with the statement according to a 7-point Likert scale (where 1 point means “disagree strongly” and 7 points means “agree strongly”). Next they answered questions about their own experiences with the type of social media that was the survey topic (e.g., podcasts). Finally we asked an open-ended question designed to reveal their bigger-picture ethical attitudes.

Naturally, the surveys covered significant ground apart from the topic of this paper; the portions of the survey that we direct our attention to here are those connected with institutional archiving of social media. Table 2 summarizes these questions. The table also recaps the scenarios, since we are aware that both the media type and the scenario details may have some bearing on the responses to general questions. For the sake of brevity, and because the questions we are analyzing here are not scenario-specific, most of the details have been omitted from the table.

Table 2. Scenarios and questions about institutional archiving on each survey

Survey media	Types of Scenarios	Scenarios and Likert scale statements about institutional archiving	Open-ended questions about institutional archiving
Twitter	Storing and reposting personal tweets and conversations; removal of libelous tweets.	*LoC is archiving public Twitter feed.* • LoC can give researchers access. • LoC can give everyone access. • LoC can give everyone access after 50 years has passed.	*none*
Photos	Storing and reposting personal photo of subjects and bystanders, including a minor; removal of photos and metadata.	*LoC is archiving public Flickr photos.* • LoC can give researchers access. • LoC can give everyone access. • LoC can give everyone access after 50 years has passed.	*none (although there are open-ended questions about actual and hypothetical photo reuse)*
Reviews	Storing and reposting reviews of an award-winning children’s book, including one that is anonymous (possibly fraudulent). Removal of reviews and comments.	*Qualified academic reviewer writes useful review of children’s book* • LoC can use review as metadata. *LoC is archiving Amazon’s book reviews.* • LoC can give researchers access. LoC can give everyone access. LoC can give everyone access after 50 years has passed. • Reviewer can remove review after 50 years has passed. • Reviews should be anonymized. • Reviews should be archived only if reviewers real name is on them.	*none (although there is a general open-ended question about reuse of content on the Internet)*
Podcasts	Storing and reposting an entertainment-oriented podcast. Removal of tags and comments. Re-use of audio snippets.	*LoC is archiving iTunes comedy podcasts.* • LoC can give researchers access. • LoC can give everyone access. • LoC can give everyone access after 50 years has passed.	Are there types of social media that the Library of Congress (or other public institutions) should not be able to archive? What are they (and why)? (*There are also open-ended questions about reuse of audio snippets.*)
Video-conference recordings	Storing and reposting of different versions of a recorded job interview, made with varying levels of consent and purpose (including satire). Reuse of video snippets, comments, and tags.	*LoC is archiving selected YouTube videos.* • LoC can archive a recording made without explicit subject consent. • LoC can archive a recording that shows only the side of a conversation where there’s consent. • LoC can archive a recording that has been satirically repurposed. • LoC can give researchers access. • LoC can give everyone access. • LoC can give everyone access after 50 years has passed.	Are there types of social media that the Library of Congress (or other public institutions) should not be able to archive? What are they (and why)?
Educational recordings	Storing and reposting of an astronaut’s commencement address with and without permission. Posting of a rebuttal from a qualified scientist. Posting, sharing and removal of reviews.	*LoC is archiving educational lectures from iTunes.* • LoC can give researchers access. • LoC can give everyone access. • LoC can give everyone access after 50 years has passed. • LoC can save reviews of the lectures without asking permission from the reviews authors. • LoC should not archive anonymous reviews.	Are there types of social media that the Library of Congress (or other public institutions) should not be able to archive? What are they (and why)?

2.3 Survey Structure

All 6 surveys included an institutional archiving scenario based on the form of social media that the survey covered. The scenario posited that a major social media company had donated its assets to the Library of Congress. Then the respondent was asked to evaluate three access variations—immediate universal access to the social media repository; deferred universal access to the social media repository; or immediate access that was limited to researchers. After we had analyzed the results to our first survey (about Twitter content), we began adding depth to these questions. Was anonymity permissible? Was permission necessary? Should comments be archived along with primary content? The most recent three surveys included an open-ended question, "Are there types of social media that the Library of Congress (or other public institutions) should not be able to archive? What are they (and why)?" Two of the earlier surveys had asked an open-ended question about respondents' reuse of online content; the answers were so interesting and provocative (see [16][17]) that we felt that respondents might express similarly diverse multi-dimensional attitudes about the institutional archiving of social media.

In the surveys, we used the Library of Congress as a proxy for any large public institution (or non-profit) capable of archiving social media. We chose the Library of Congress for three reasons. First, respondents may be familiar with the Library's Twitter effort; it makes sense to extend this project to other types of social media. Second, it seemed necessary to select an institution with sufficient resources and expertise to make the scenario plausible. We could have used the Internet Archive, but respondents were slightly less likely to be familiar with its successful efforts to date. Finally, we wanted to use a public governmental institution so there would be no overriding questions about the scenario's legality; it is easy for respondents to get distracted by peripheral aspects of a scenario.

3. RESULTS

The surveys elicited concerns about institutional archiving of social media content in two ways: Likert-scale responses to specific statements about the scenarios, which yielded quantitative data, and an open-ended question at the end of the survey that asked respondents what they felt should be excluded from a hypothetical Library of Congress social media archive. We first discuss the survey-specific quantitative results; these responses are all in the context of our media type-based scenarios.

3.1 Media Type-Specific Results

We present the media-specific results from four new surveys on reviews, educational recordings, podcasts, and videoconferences, so the results can be compared to our survey results for tweets [16], photos [17], and build on an earlier paper comparing the simple institutional archiving scenarios for those two types [18].

Amazon book reviews. The third survey of the series posed scenarios that involved reviews of the classic children's book, *Where the Wild Things Are*; the reviews—including one written by an expert on early childhood education, and one purportedly written by a child—were published on Amazon's online bookstore, and had amassed comments and ratings. The survey also included our standard institutional archiving scenario: this time, the Library of Congress archive was using the Amazon reviews as metadata to help describe books in its collection. We asked respondents about how they felt access to this archive of reviews should be regulated: should it be publicly available, available to researchers only, or publicly available after a significant period of time (50 years) had elapsed.

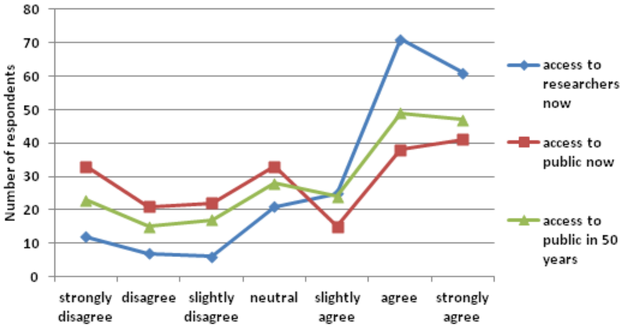

Figure 1. Three access conditions for archive of reviews

Figure 1 shows respondents' relative attitudes toward the three access conditions; the differences between them are statistically significant (Wilcoxon's Signed Rank, p<.001 for closest). Current access by researchers elicits the most positive reaction; current access by the public elicits the least positive reaction and the most controversy (as evidenced by the bimodal distribution of responses). Deferring public access for 50 years is both less controversial than immediate access, and less well-received than limiting access to researchers.

We tested the concept of permission with two questions. In one, the review on Amazon's website was written by a child (who is anonymous), and in the other, the review was written by an adult, a credentialed scholar in early childhood education who is authenticated by the system. Is permission necessary before the Library of Congress ingests either review into its repository? The status of the review (and reviewer) does not seem to affect respondents' attitudes toward permission. The concept is controversial, with respondents just about evenly split in both cases. The means are 3.98 and 3.93 (4 is neutral) for the child's review and the scholar's review respectively, and the distributions are flat.

Educational recording results. The final three surveys covered different types of recorded media. First we will examine the results for the scenario about institutional archiving of educational recordings, such as those found in iTunes University. To set the stage, the scenarios earlier in the survey posited a recorded commencement lecture delivered by an astronaut and a recorded rebuttal posted by a noted geologist. As Table 2 notes, we also tested concepts of permission and anonymity in this survey; these results are discussed when we compare results across surveys.

Responses to our standard access probes—access to researchers now, access to the public now, and access to the public in 50 years—revealed that the public access conditions (now and in 50 years) were statistically indistinguishable. However, access to researchers now tested more positively than either public access scenario (with statistical significance, Wilcoxon Signed Rank, .05>p>.02 for both).

Podcast results. Next we look at the institutional archiving results for the podcast scenario (the scenario that posited entertainment-oriented podcasts donated to the public institution by an iTunes-like service). Again, granting immediate access to researchers was significantly different than the responses to the two public access scenarios (Wilcoxon Signed Rank, p<.001 for both). There was no significant difference between the two public access scenarios.

Videoconference results. Finally we look at the institutional archiving results for the videoconference scenarios (several scenarios that posited recordings of job interviews that were manipulated and published with a variety of purposes in mind, including instruction, satire, and vlog-like personal commentary). Given popular YouTube videos as the archival genre, we found no significant difference between the three access scenarios. This result may have been because the videoconference recording survey attracted fewer respondents. Nonetheless, access to the material looks generally uncontroversial and trends positive.

In addition to testing our standard access conditions, we also explored the institutional archiving of a videoconference given varying levels of consent (to being recorded) and anonymity. In one scenario, a job interview conducted over Skype is recorded by the interviewee (Bill) without explicit consent of the interviewer (or his company). In the second scenario, the interviewer's side of the conversation has been removed from the final video, as well as any traces of corporate identity. In the third case, the recording has been satirically repurposed; again, the interviewer's side of the conversation has been replaced on the video, and any traces of corporate identity have been removed. Respondents were asked to evaluate whether it was appropriate to archive each of the three versions of the videoconference, and whether anonymity would influence this decision (i.e. if consent could not be obtained from the video's creator). Figure 2 shows the results of these conceptual probes about consent, permission, and anonymity.

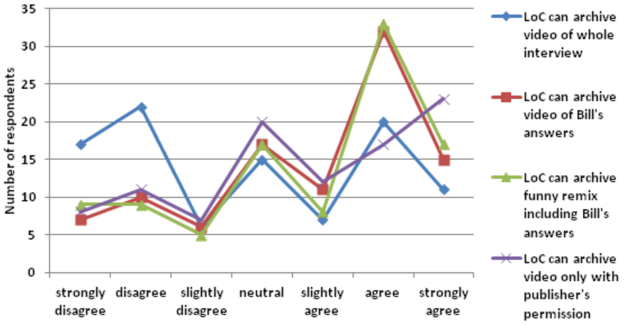

Figure 2. Conceptual probes of consent, permission, and anonymity

The results show that respondents were skeptical about the Library of Congress's right to archive the video of the whole interview. Both of the edited versions fared better; respondents assessed the institutional archiving effort positively. The differences between the full video and the two edited videos are significant (Wilcoxon's Signed Rank, p<.001 for both). There is no significant difference in the responses for the two edited versions of the video; in other words, the differing genres and purposes of the videos (instruction in one case and satire in the other) did not appear to influence respondents' sense of whether the video could be archived by a public institution.

A final condition—whether the Library of Congress could archive a video only if it was able to get the creator's consent—elicited similar responses to two edited video conditions, albeit slightly less enthusiastic. We should note that in the other cases, Bill's consent was only implied because he made the video public on websites like YouTube and FunnyOrDie; his permission was never explicitly solicited. This seems consistent with the response to the open-ended question we included at the end of the survey (and discussed in a later subsection)—some respondents feel that publishing content implies this sort of consent.

3.2 Cross-Media Results

In our past research, we used the Library of Congress scenario (the Library of Congress is building an archive of the media type used in the survey's other scenarios) to test three conditions (immediate public access, delayed public access, and immediate researcher access). In particular, we compared responses to the Twitter survey and the Flickr/personal photo survey. We found that the responses seemed to vary across media types [17]. How would this finding play out over six media types and genres?

Researcher access now. Unsurprisingly, the most positive responses are elicited by researcher access to an institutional archive that consists of the four most 'published' media types and genres—educational content comes out on top, followed closely by reviews and podcasts, with YouTube videos slightly below. Educational recordings are also the least controversial.

Which media types are controversial, given this scenario of immediate researcher access? More personal (and possibly less valuable) material proved to be more controversial: tweets and photos. Figure 3 shows this ordering.

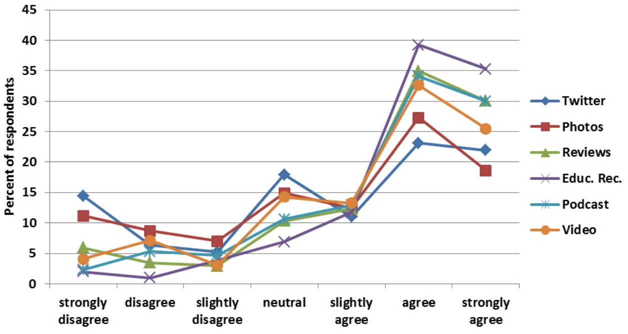

Figure 3. Immediate researcher access by media type

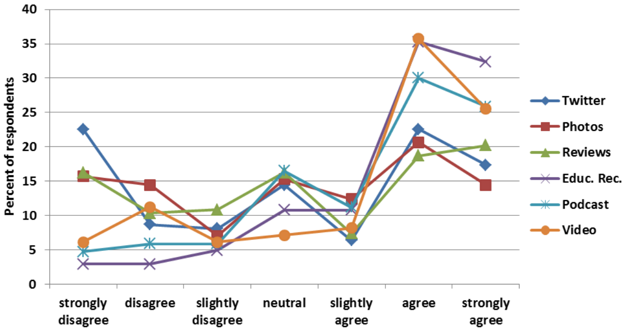

Figure 4. Immediate public access by media type

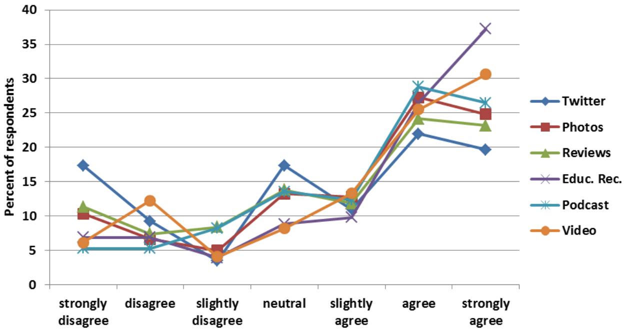

Figure 5. Public access in 50 years by media type

Public access now. Given the broadened purpose of public access to these media type-based institutional archives, we might expect positive outcomes to remain the same and negative responses to be exaggerated. We might also expect a heightened sense of controversy. Indeed, Figure 4 shows the responses to immediate public access to educational recordings, podcasts, and popular videos to trend positively, and the respondents to be in rough agreement.

On the other hand, public access to tweets, photos, and product reviews elicit less agreement, with tweets being the most controversial type. What might cause this controversy? Tweets and photos may rely on 'privacy through obscurity', but surely reviews rely less on this effect. Later, through the open-ended responses, we may speculate that the controversy may be engendered by the respondents' sense of the ultimate value (or lack of value) of this content.

Public access in 50 years. Does the passage of time mitigate privacy concerns? If it did, tweets might no longer be controversial relative to the other media types. Figure 5 shows that this reordering does not occur. Instead tweets remain controversial (and educational recordings remain benign). As we saw in our earlier comparison across types [17], concerns about access to photos are indeed diminished by waiting 50 years to provide access to them. Yet access to popular videos (across video genres) shows continuing concern. These effects again may be explained in the responses to the opened ended questions—characteristics such as social value may be amplified with time, especially as the perceived expense of maintaining an archive grows.

Anonymity. One of the features we were interested in exploring was anonymity. User-contributed content sometimes relies on relinquishing anonymity; some social media sites demand 'real user' authentication. Is content valuable if its creator cannot be identified? Might anonymity mitigate some of the privacy loss that individuals experience when social media is archived? We included some questions about creator anonymity in our institutional archiving scenarios.

Figure 6 aggregates the responses to three related questions about anonymity. First, respondents reacted in a positive way to the requirement that the Library of Congress anonymize book (or product) reviews it archives from Amazon by removing the name or pseudonym of the reviewer. We speculate that respondents feel this may address concerns about privacy loss. The two other results confirm this finding about reviewer anonymity; respondents were ambivalent about restricting the ability of the Library of Congress to archive anonymous reviews of educational recordings. Furthermore, respondents tended to disagree with limits placed on the Library of Congress's efforts to archive Amazon book reviews to only those with real names attached.

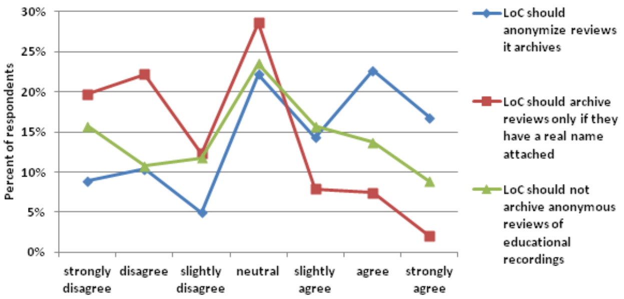

Figure 6. Exploring creator anonymity

These results also underscore the prevailing current of thought among respondents that a public institution should have an overriding ability to archive material for the public good. We will discuss this perspective in the next section.

Archiving associated content. One important property of social media—be it user-contributed photos, videos, reviews, or recordings—is that it may have associated content such as comments, ratings, tags, and other annotations that add value to the original item. Which restrictions should be imposed on archiving this additional content?

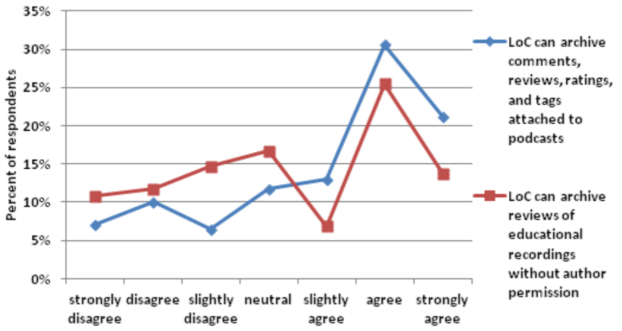

Figure 7. Archiving associated content

Figure 7 aggregates responses to two representative questions about whether the Library of Congress can archive this type of social metadata, with and without author permission. In the podcast survey, respondents were asked to evaluate whether the Library of Congress can archive the reviews, ratings, and tags that listeners associate with a podcast. They generally agreed that social metadata can be ingested along with the primary media. When we added the caveat that the social metadata would be ingested without asking the authors' permission (when educational recordings were archived), the effort became more controversial. In fact, many respondents raised the idea of creator permission when they answered the open-ended question described in the next section, possibly without realizing that social metadata contributors often cede their rights to the content as per the terms and conditions of many social media services.

3.3 Exclusion

The final open-ended question of the three recorded media surveys asked respondents to consider the general problem of archiving social media. Should the Library of Congress simply be able to archive any social media that's available today on the public Internet? If not, what should be excluded and why?

The surveys were lengthy, yet most of the respondents gave this question some thought; they seemed anxious to express their own opinions about whether institutions should be archiving social media as part of their mission, or whether this sort of archiving was an opportunistic incursion into individuals' private lives.

How did respondents reason about this question? How did they determine what should be excluded? Some (about 1/3 of the respondents) considered the purpose and mission of the Library of Congress as a public institution, often appealing to an aspirational notion of social good and everyday culture (although a few respondents, as we will discuss, did not think this sort of archiving fell within the Library's purview). Other respondents reasoned from the creator's perspective: about 1/5 thought that surely the Library should be obtaining permission from the media's creator. Still others considered the value—and, potentially, the veracity—of the material: is it worth it to gather seemingly trivial cultural detritus? A small number (about 5%) noted that the technology itself offers remedies to this problem: privacy settings and other mechanisms telegraph the individual's intent; these respondents thought the Library of Congress should use existing technological affordances to guide their efforts. Finally, about 10% of the respondents situated their answers in the legal concept of public domain, primarily indicating their belief that the Internet affords no copyright protection. We examine each of these stances in turn.

We began our analysis by open-coding the responses, attempting to capture the theme of each respondent's primary argument (and how it differed from the arguments other respondents had offered) [24]. Many of the respondents reasoned from several different vantage points, creating a nuanced set of exceptions and edge cases; we try to indicate a rough sense of how prevalent various perspectives were in the arguments. Generally, we only report on a particular point of view if it occurs in a meaningful number of responses (more than 1%); this way, we are able to indicate if a response is a true outlier. We identify the respondents according the survey (EDU, PC, or VC) and a respondent-specific ID number. Thus if the ID number matches, e.g. EDU007 and PC007, it means that Respondent 007 completed both the EDU and PC surveys. Duplicate responses from a single person completing different surveys are only counted once in computing the relative prominence of a response type.

3.3.1 The Creator's Perspective

It is not difficult to understand why many respondents took the creator's perspective, especially given the current climate of anti-piracy legislation, and the controversy surrounding the Library of Congress's efforts to archive the public Twitter feed. Respondents brought up two main facets of this perspective: permission and the implicit privacy of personal content, especially when it is emotionally resonant or may be damaging to the individual. Needless to say, there are many variations of each type of response; we explore only the main ones.

Permission. The need to obtain creator permission arose in about 20% of the responses. One respondent argued, "*The Library of Congress should not be able to archive personal imagery or opinionated media (audio, video, textual or otherwise) without the given consent of its owners. This is because everything created deserves to have proper credit by its creator, and if the creator doesn't want it archived that should be an available choice. The Library of Congress doesn't automatically own something just because they want it.*" [PC007]

Naturally respondents identified exceptions. Several respondents brought up mitigating factors such as the prevention of pedophilia or national security needs. A few also inserted consent earlier into the publication process: *"Maybe, if the media company supplies an option to opt out of future archiving then it should be honored.*" [EDU187] or later "*The Library of Congress should be able to archive anything that is freely available on the internet, but should provide the ability for the content creator to remove the content from the archive.*" [PC074] Currently this consent is part of many service providers' terms and conditions, but respondents remind us that institutional archiving is not something that's on their minds, even if they've read the fine print.

Anonymity is the flip side of permission: if a contributor does not dignify content with attribution, why would a public institution archive it? "*Anonymous videos or videos without consent should not be archived, as they violate freedom of speech.*" [VC005] Such arguments were relatively rare (only 4 occurred in the 3 surveys); these respondents seemed to be drawing on encounters with abusive anonymous comments on YouTube or blogs, and several also may have misunderstood 'freedom of speech.'

Privacy. About 1 in 7 respondents express a desire to protect the creator's privacy. These respondents argued that personal material shared as social media is *per se* private: "*Facebook pages, Twitter accounts, and any other social media on a personal level should not be archived. It is that person's private account and should not be messed with.*" [PC020] About 1 in 10 argued further that some social media types—e.g video, personal correspondence, or Facebook posts—are more private than others: "*…Videos are a lot more personal than anything written on paper so they should be treated more cautiously.*" [PC245]

A common theme running through privacy arguments—one that echoes Stutzman's example—is that this personal material is intended to be transient. That this transient material may become permanent is disquieting to about 3% of respondents, and to some, seemingly unfair: *"...It just seems wrong that something made when you are 15 would follow you forever"* [VC114]

Similarly, a few respondents pointed out that comments on social media were created in a context—in open discussion at a particular time, in a particular forum—and preserving them in an archive would serve to decontextualize them in such a way that is unfair to the creator: *"...Even though the comments are public, people making them think they will remain on that site and for the site members, not be archived."* [PC107]

Other creator-derived concerns. In addition to the need to obtain the creator's permission and a desire to protect the creator's privacy, a few respondents brought up other concerns such as the subjects' privacy (drawn from photography) and the concomitant need to seek the subjects' permission. Respondents also raised the potential need to protect the contributor's interests if the content is commercially valuable.

3.3.2 The Content Perspective

If a public institution is to invest significant resources in an archiving effort, to many it follows that the effort should be worthwhile; the content should be sufficiently valuable to merit attention or, from the standpoint of protecting society, it should not perpetrate harm. In this case, the content's creator or source is not a factor; rather, the content speaks for itself. We explore elements of the two major content perceptions: the content should be intrinsically valuable and the content should cause no harm.

High-value content. Some respondents (under 10%) thought in terms of inherent value to society, while others posed examples of content or content genres worthy of archival efforts. For example, PC153 argued that the Library of Congress should not archive "*Anything that is not considered public record, or that does not provide some inherent value to society. No one cares what status updates someone in Colorado writes about the sandwich they ate for lunch*." Others kept their responses generic. For example, EDU078 responded, "*I think if [social media content] could benefit someone else, then yes, [it] should be archived."*

Because the scenarios focused on recorded material, some respondents thought in terms of recorded genres: *"If the Library of Congress deems something is worth archiving, I believe they should be able use it. Things only available online, perhaps even youtube. If the recordings offer educational opportunities for the future posterity, I think its [sic] a good thing.*" [EDU046] Note the emphasis on unique content: EDU046 is consciously excluding material that has migrated from published physical media—the respondent may be thinking of movie clips, commercials, or excerpts from published forms like movies or TV shows. Another respondent's reaction demonstrates a lack of understanding about the value of archives, "*Youtube is so popular there would be no point in archiving any of the videos that are already available to public unless it pertains to specific things that may be helpful.*" [VC048]

Some respondents did not feel all public content should be archived; specific types were excluded as *per se* valueless. Tweets and YouTube videos were particular targets of this reasoning, e.g. "*Although things like Twitter are generally public, I think the Library of Congress archiving individuals' accounts is tacky and in poor taste. A podcast or an essay is different than a tweet.*" [PC166] and "*Personal videos because I find it excessive and unnecessary.*" [VC083]

Others argued that the content of social media services such as Facebook is trivial, or that the few nuggets are obscured by the weight of the mundane, e.g.: "*People who really post worthwhile information usually don't bother with facebook. Of course there may be exceptions, and I believe that those should be taken into account, but as a general rule it is not worthwhile to comb through millions of facebook posts just for the few good ones."* [PC025]

Thus arguments about social media content value frequently shifted the ground to what might be historically important (and what would make it so, say Obama's tweets). Here the respondent sometimes considered the content creator as a property of the content, not as a stakeholder in the archival process, e.g.: *"...social media from the average citizen [should only be saved] if the person is a political power or somehow of historical significance (IE President of the nation, a nobel prize winner etc). Otherwise I think recording the average persons tweets etc would just be a colossal waste of time.*" [EDU035]

Content veracity. Content veracity was tied to value in a little over 7% of the responses. Opinions, these respondents opined, are not worth the bits they are stored as: "*I don't think opinions should be archived if they are not useful to future generations.... This is the reason I don't read blogs. If a person does not have sound experience and knowledge in the topic being presented (say a podcast for example), then I have no interest in wasting my time reading about how they 'feel' about said podcast.*" [PC008]

Do no harm. Content veracity also formed the linchpin to many arguments about the harm done by social media. Inaccurate content, about 6% of respondents argued, is inherently pernicious. "*Personal profiles [on Facebook] can be a deception.*" [PC131] We might speculate that either bad experiences with inaccurate information, or accounts of such experiences, led some respondents to limit archival information to that which has been verified: "*I'm not sure why the Library of Congress would be involved in social media. ... I think the Library of Congress is more useful by keeping information that has reasonably been researched to know is true, as some of these things out there are not real.*" [EDU180] Using the same logic, several respondents argued against keeping anything that is incomplete.

Just as inaccurate information may be regarded as pernicious, so too is material that is considered either biased, slanderous, offensive, racially-charged, or would incite lawless behavior, e.g. "*I think social media that contains racial bias or any kind of prejudiced based content shouldn't be archived.*" [PC043] and "*…Items that could incite lawlessness, promote illegal activities or provide information that could be used detrimentally against the welfare of others may be candidates for information that should be withheld.*" [EDU040]

Pornography was called out as a particular type of social media content that shouldn't be archived by a public institution (social media sites themselves seem to expend considerable effort on anti-pornography policies). Some respondents were concerned that no offensive material be accidentally (or purposefully) harvested in the name of social media preservation, e.g., "*[no] Porn, off-color, anything that could harm a child or person if shown [should be included in social media archive].*" [VC024]

Again, many of these perceptions seem misaligned with the purpose of archives, or perhaps confounding the idea of an archive with the notion of a reference like Wikipedia.

3.3.3 The Technology Perspective

Social media services usually include mechanisms such as user authorization and privacy settings designed to protect their users' interests. Some sites (like Facebook) even do a certain amount of policing to encourage their users to stick with their real identities, and not to adopt personas based on fictitious characters (e.g. Nomi Malone from *Showgirls*), so the authorization is a meaningful association of a real person with the content they contribute.

Only 5% of the responses suggested that institutional archiving efforts should simply rely on authorization mechanisms and privacy settings to guide collection policies, e.g.: "*[Institutional archives should exclude] social media that has expressly been blocked by the user, such as Twitter profiles that are kept private. These types of profiles indicate the user's express wish not to have their information accessible to everyone.*" [PC246] Another response revealed concern about 'friends of friends' and the uncontrolled (and, practically speaking, uncontrollable) flow of personal information through these services:

> "*The Library of Congress should not be able to archive any types of social media content that is accessed through an authorized user but not available to the general public (i.e. Facebook, Myspace). I believe that information which users knowingly make available to everyone is fair game, but the line should be drawn when content is distributed to the Library of Congress by individuals with access to information not available to everyone. A user's friend on Facebook, for example, should not be able to share information not intended for the general public with the Library of Congress without the original poster's permission.*" [EDU154]

The permeability of privacy settings and other mechanisms to control personal information seemed to convince respondents that these mechanisms provide insufficient guidance to determine which content is intentionally public.

3.3.4 The Legal Perspective

As we saw in earlier studies, a significant portion of respondents consider material posted on the Internet to be in the public domain; around 10% explicitly say so, and as many as 1/3 imply it. In this case, they may feel that anything public should have no restrictions, especially in view of the Library of Congress's role as a governmental institution. We divide these arguments into two (uneven) camps: the small number of respondents (under 3%) who feel that the Library of Congress is charged with maintaining the public record, and the much larger number who feel that anything on the Internet that is publicly accessible falls into the public domain.

Public record. Public record arguments have at their heart the idea that institutions like the Library of Congress are free to gather published (and public) information as they see fit, as per their defined role as archival institutions, and that this overrides interests of the individual. In other words, the Library of Congress is serving a role which gives it special dispensation to curate the public record. PC059 maintained that archiving should proceed without restrictions: "*Though the LoC is performing these actions with public fund, people should consider that anything digitally recorded and uploaded to the internet may exist outside their control "forever*". Similarly, PC170 said, "*…the Library of*

Congress should be able to archive any public information they would like. Due to public information status." PC258 simply declared, "*I feel all social media is public record*." Finally, VC231 likened YouTube videos to print books: "*Once the video is published for all to see, then the Library of Congress should treat videos as they treat printed material*."

Alternatively, a small number of respondents (around 3%) were concerned about how copyright law applied to this type of endeavor. In so doing, they often revealed a fairly loose grasp of copyright and its purpose: "*They should not be able to archive media, like music, that is copyrighted because it is not in their jurisdiction*." [PC081] There is also a perception that the Library of Congress serves a regulatory function, and that the materials it gathers are some sort of yardstick for measuring the truth (this goes hand in hand with our earlier discussion that this type of archive should somehow be factual and accurate): "*I think it is important for the Lobrary [sic] of Congress to regulate material.*" [PC244] One respondent even felt a social media archive would have an evidentiary role: "*they can go to their archive files for evidence*." [PC095] It is not clear who 'they' refers to, although several respondents took 'they' to mean Congress (in keeping with history, whether by intention or accident).

Public domain. On the flip side of this public record argument, other respondents reasoned that because some social media is part of the public Internet, the content is all in the public domain. Why should restrictions be imposed on the Library of Congress that are beyond those applied to everyday users? "*Once something is on the Web, it belongs to the Web users*," wrote VC141. VC169 explained, "*I think if someone puts something publicly on the internet, they have no control over how it is used in future. Therefore, anything that an individual chooses to put on the public domain is in fact public. If someone does not want something archived they should keep it in the private domain.*" Public domain, private domain: if the content is accessible to any of us, it is ours.

3.3.5 Social Good

Although many respondents implicitly considered the role of the Library of Congress in their reasoning about what should be in a social media archive, some evoked it explicitly. What are the intentions of a public institution when it gathers material that is potentially private, that is often personal, and that might be damaging to individuals? "*They should be able to [archive social media] if they have good intentions as far as what they will do with that info…*" [PC112]

Around 10% of the respondents argued that the Library of Congress should be able to archive everything as a matter of principle ("*If they archive some they should get them all*!" [EDU156]), either because it all has documentary importance ("*… it will all be useful for research*." [PC049]) or because taken as a whole, social media content has unique cultural value ("*…future generations can use all types of social media to learn about our current time*." [PC071]). However, respondents also acknowledged counterarguments about content value, e.g.:

> "*If given the space, most things should be documented. However trivial it may be*." [PC030]
>
> "*…Podcasts... interviews... even goofy user videos are all part of society/culture and worth archiving for future accessibility*." [PC157]
>
> "*The Library of Congress should really keep at least a sample of everything. We are creating culture and history. No matter how some people feel about a certain subject or genre, nothing should be excluded*." [PC181]
>
> "*The Library of Congress is charged with (among other things), recording our history. History is made by citizens every day and social media is a good way to capture this history*." [PC193]

One way to mitigate perceived damage is to separate preservation and access. Several respondents made this separation explicit—that not everyone should have access to the archive—and wrote that the data should be "*well protected*." [PC112]

Naturally in the current political climate there will be nay-sayers about the value and mission of governmental institutions. Two responses questioned not just the effort's Herculean nature, but also whether archival institutions should exist and be involved in the affairs of individuals. "*I don't believe the LoC should exist as it's not a private institution.*" [PC117] Another respondent restored the responsibility to the individual (without considering that the archive would fulfill a larger role than keeping personal material safe for oneself): "*I feel that this should be a person's own responsibility and that the government should have nothing to do with this… There are plenty of private digital vaults around for anyone to avail themselves of, if they wished*." [PC255]

4. DISCUSSION

Personal experience with social media, coupled with a growing unease about larger issues such as privacy, piracy, the control of user contributed content, and the ubiquity of technology in modern life, has greatly complicated institutional efforts to archive social media. [11] By fielding a set of six surveys, we have been able to paint a more complete and nuanced picture of attitudes toward current and future efforts.

First, we have learned that content type matters: different content types evoke different responses when we propose that they be saved on a grand scale for posterity. Two dimensions guide and constrain how respondents react to different content types:

(1) How personal is the genre or type perceived to be? On one hand, the more personal (and seemingly transient) the information is, the more it raises the threat of additional privacy loss, and on the other hand, the less certain respondents are about its long-term value to society. We have also seen that not all elements of a social media genre are equal. Comments, for example, may seem more private than the primary media.

(2) How familiar is the genre or type? How much experience do respondents have with it in their everyday lives? Have they created it as well as simply consumed it? In our photo survey, we learned that respondents were well able to rationalize individual instances of reuse when they understood the motivation and had engaged in it themselves [17]. In this set of surveys, we have become increasingly convinced that it is familiarity with the creation and reuse of a genre or type that leads respondents to come up with plausible examples to reason from. Furthermore, the scenarios that we use to set the stage seem to have a substantial effect on respondents' attitudes to more general questions about institutional archiving.

Next, we have seen a need to tease apart collection-building and access. These distinctions do not occur to respondents naturally. When will a collection be used, by whom, and for what? Certain limitations on access would seem to ease anxieties about gathering, ingesting, and processing the content in the first place.

"*This is just too hard of a problem*" complained one respondent. The idea of this survey was not to throw the question back onto individuals, but rather to help institutions frame archiving efforts so that boundaries match the public's expectations. What are the features that people care about? For example, the idea of harm (either emotional or professional) arises when respondents imagine their social media content taken out of context; that they are ambivalent about the preservation of hate speech seems a reflection of larger cultural impetuses, but not necessarily aligned with the mission of any archiving effort.

In the end, the surveys highlight most profoundly the need to help people envision boundaries—collection boundaries, access boundaries, issues of identity and attribution, of permission and reuse. Most importantly, characterizing current attitudes will help us determine if these attitudes are changing with time and experience, and whether they are tied to demographic properties.

In practice, precautions like obtaining permission are not straightforward. The Preserving Virtual Worlds project noted that its efforts to obtain in-world permission were fraught with difficulty; they cite an at-best permission rate of 10% and at worst their efforts were met with hostility [19]. This reaction seems to stem from peoples' problems envisioning constructive reuse. Although constructive reuse by artists [13] and scholars [22] is within reach, our study suggests that respondents' imaginations about everyday reuse are shaped and constrained by their own experiences.

Broader still—extending beyond institutional archiving—is the perspective that much of this new content is viewed by its creators and consumers as largely personal, even as it is increasingly entangled in relationships with service providers that render it more closely akin to published content. Just as the web caused us to rethink the rhythm of fixity and fluidity [14], so too has social media caused us to reconsider the relationship between personal, shared, public, and published media.

5. ACKNOWLEDGEMENTS

This work was supported in part by National Science Foundation grant DUE-0938074.

6. REFERENCES

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

[2] Boyle, J. *The Public Domain: Enclosing the Commons of the Mind*, Yale University Press, New Haven & London, 2008.

[3] Connelly K, Khalil A, Liu Y. "Do I Do What I Say?: Observed Versus Stated Privacy Preferences," *Proc. INTERACT 2007*.

[4] Downs, J., Holbrook, M., Sheng, S., and Cranor, L. Are your participants gaming the system?: Screening Mechanical Turk workers. *Proc. of CHI'10*. ACM, 2399-2402.

[5] Faridani, S., Bitton, E., Ryokai, K., and Goldberg, K. 2010. Opinion space: a scalable tool for browsing online comments. *Proc. of CHI '10*. ACM, 1175-1184.

[6] Gilbert, E. and Karahalios, K. Understanding Deja Reviewers. *Proc. of CSCW '10*, 225-228.

[7] Hill, B., Monroy-Hernandez, A., and Olson, K. Responses to Remixing on a Social Media Website. *Proc. AAAI Conference on Weblogs and Social Media*. 2010. 74-81.

[8] Ipeirotis, P. Analyzing the Amazon Mechanical Turk Marketplace. *ACM XRDS 17*, 2, Winter 2010.

[9] Jakobsson, M. (2009) Experimenting on Mechanical Turk: 5 How Tos. *ITWorld*, September 3, 2009.

[10] Java, A., Song, X., Finin, T. and Tseng, B. Why We Twitter: Understanding Microblogging Usage and Communities, *Proc. KDD'07*, ACM, 56-65.

[11] John, J.L., Rowlands, I., Williams, P., Dean, K. *Digital Lives >> An Initial Synthesis version 0.2*. Digital Lives Research Paper Series, British Library, 3 March 2010.

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

[13] Lessig, L. *Remix: Making Art and Commerce Thrive in the Hybrid Economy*, Penguin, New York, 2008.

[14] Levy, D. "Fixed or Fluid: Document Stability and New Media." *Proc. ECHT'94*, 1994, 24-31.

[15] Library of Congress: Digital Preservation Video Series. 2010. *Digital Natives Explore Digital Preservation*.

[16] Marshall, C.C., and Shipman, F.M. 2011. Social media ownership: using Twitter as a window onto current attitudes and beliefs. *Proc. of CHI'11*. ACM, 1081-1090.

[17] Marshall, C.C., and Shipman, F.M. 2011. The ownership and reuse of visual media. *Proc. of JCDL'11*. ACM, 157-166.

[18] Marshall, C.C., and Shipman, F.M. 2011. Attitudes about Institutional Archiving of Social Media. Proc Archiving'11.

[19] McDonough, J., Olendorf, R., Kirschenbaum, M., Kraus, K., Reside, D., Donahue, R., Phelps, A., Egert, C., Lowood, H., Rojo, S. *Preserving Virtual Worlds Final Report*. 8/ 31/2010.

[20] Munson, S., Avrahami, D., Consolvo, S., Fogarty, J., Friedman, B., Smith, I. Attitudes toward online availability of US public records. dg.o '11, ACM Press, 2-9.

[21] Odom, W., Sellen, A., Harper, R., Thereska, E. Lost in Translation: Understanding the Possession of Digital Things in the Cloud. to appear in *Proc CHI 2012*.

[22] Owens, T. Tripadvisor rates Einstein: using the social web to unpack the public meanings of a cultural heritage site. *IJWC 8*, 1, 40 – 56.

[23] Shinen, B. Twitterlogical: The Misunderstandings of Ownership. http://www.canyoucopyrightatweet.com/

[24] Strauss, A. and Corbin, J. *Basics of Qualitative Research*, Sage Publications, 1998.

[25] Stutzman, F. Twitter and the Library of Congress. http://fstutzman.com/2010/04/14/twitter-and-the-library-of-congress/.