Social media, personal data, and reusing our digital legacy

Catherine C. Marshall
Microsoft Research, Silicon Valley

I remember what the Web was like in 1994. That's when I put up my personal homepage and when many of my colleagues published theirs too. The initial rush of personal homepages joined the physics preprints, the websites constructed with a hobbyist's fervor (for example, documenting the Klingon language, early music, or San Francisco graffiti), and the first few web-based businesses (such as the online wine retailer Virtual Vineyards and the online magazine *Hotwired*). The fight for readers' attention had not yet reached a fever pitch, although a few scholars like Richard Lanham (1993) and David Levy (2001) anticipated that it would, picking up on a recurring century-old leitmotif of attention as the scarcest resource (Thorngate, 1987; James, 1890). In short, it was still a human-scale hand-made Web populated at least in part by navigable user-contributed content. Even official content (e.g. the city homepage for Bremond, Texas, "Home of Friendly People & Polish Sausage"[1]) was often the product of a single enthusiast who'd hand-edited the HTML.

Local storage was tractable as well: most of us had a work computer and a home computer. Digital cameras were newly on the market. Doing anything significant with video required an Avid editing bay, a proposition too expensive and too technically difficult for the average consumer. Computer Science graduate students played Doom, a shoot-em-up video game that was laughably primitive by today's production-heavy standards; there were no characters to level, no voice-over actors doing the vocal sound effects—the oofs and grunts of effort and pain—and there was no long-term investment in the game. Email was largely a simple affair; the occasional MIME type attachment was regarded as a nuisance and a challenge to open. In short, you probably had a manageable number of files you cared about, which were stored in a fairly small number of places.

At that time, many of our personal information management problems—and hence our personal archiving problems—stemmed from the mix of digital and paper files in our lives (Sellen and Harper, 2001). Writing this chapter, for example, might mean crossing from digital to paper and back again many times; maintaining a history of the chapter's versions would involve coordinating between alternating print documents and digital files.

Soon corporations and commercial concerns seized upon the Web. Bureaucracy moved online. And for a while, the Web became the province of professional designers and media creators. Creative types moved to places like New York and San Francisco, where tiny startups and giant corporations hustled and competed to enlist what came to be called content creators. Much of what we used on the Web wasn't ours, and we had little apparent interest in keeping it safe.

[1] This page can be retrieved from the Wayback Machine. http://www.rtis.com/reg/bremond/, version 13 February, 1997. This version of the page from 1997 agrees with my memory of what the city's homepage was like in late 1994, when I first encountered it.

But soon thereafter, user contributed content again came to play a significant role in life online. Naturally this shift did not go unnoticed; writing was framed in terms of addressing the interests of a long tail of readers (Anderson, 2004), and curation was put into the hands of the crowd, who would ultimately be wiser than editors (Surowiecki, 2004). Social networking sites like MySpace (then Facebook) picked up momentum. At the same time, everyone had more ways to record things, more ways to edit things, and more venues to share what they had.

Thus we have arrived in 2013 with three puzzling pieces of baggage left over from an earlier era of computing. These issues are essential to the way we think about personal digital archiving.

1. The first is long-term keeping under the guise of *backup*: We have long thought it advisable and necessary to back up the precious digital files on our computers to keep them safe; so why is it so hard to get people (including me) to back up their files?

2. The second extends backup to our stuff *on the cloud and in social media*: should we think of our stuff on the cloud and in social media as an extension of our local stuff and does this mean we should have a plan for keeping it safe too?

3. Finally, the third stems from the complexities of *ownership and control* that accompany storing stuff on social media services: Is that online stuff still under our control? What do we own and what can we use? Does it have value to other people?

All three of these issues represent rational practices and ideas: Who really wants to lose stuff? Why would the stuff on social media be different from the stuff we store locally? Who wants to shrug off the protections afforded by existing notions of ownership and control? But at the same time, none of the old ways of doing business make sense the way they used to.

Sometimes I log on to my old account at Texas A&M University, where I was a research faculty member in the mid-1990s. Besides the files that I use for my personal website, everything else is just as I left it. And compared with the files I've created since, there's almost nothing there. At the time, these files (largely research data and publications), plus some short stories and personal email that I had stored at home made up the whole of my digital belongings in 1995. That it now seems so spare and so localized speaks volumes.

The locus of personal information and people's associated management practices have shifted dramatically over the past decade. I have seen this not only through reflecting on my own practices (which is dangerously limited, as I will demonstrate, as is reflecting on the practices of one's own social network—you must get out there to see what is going on), but also through the window of a series of studies. The studies use a variety of methods—some are slanted toward detailed snapshots of individual use and others take a broader (but necessarily shallower) perspective. They are all looking at personal information from the vantage point of the technologies we use to keep it and the practices that we use to control it.

One way I like to look at the study results is to single out what surprised me at the time. I often walk into a study thinking one thing, and I come away from data collection and analysis thinking something entirely different. In this chapter, I'll quickly discuss three studies performed over about seven years and tell you what surprised me in each. Don't think of them as studies for their own sake, but rather as signifiers of larger changes that were afoot. Even the questions that it made sense to ask have changed.

Study 1. Stumbling on curation via benign neglect

Study 1, performed in 2005 with my colleagues Sara Bly and Francoise Brun-Cottan, was the first one we had done that specifically took on the question of digital archiving (Marshall, Bly, and Brun-Cottan, 2006). We wondered how people were trying to keep their digital stuff safe over the long haul, and whether they were bringing physical metaphors to the problem. We used a professional recruiter to find study participants in three West Coast cities, reasoning that if we stuck to the San Francisco Bay Area, we might get too many technical people. We required that participants use a computer and a varying number of other devices like digital cameras and recorders, iPods, and PDAs. (Smart phones had not yet been introduced). So the people we interviewed for the study were diverse—a performance artist, a few high school and college students, a young auteur, a woman who ran a day care center, an urban beekeeper, a real estate broker and mom, a young single professional guy, a blue collar worker in a lumberyard who was a NASCAR fan, and a psychotherapist. Indeed, at first blush they seemed to have little in common.

But after we'd interviewed them, it turned out they did have something in common: they had islands of deep understanding (for example, several could use Photoshop in sophisticated ways), but these islands existed amid roiling seas of confusion. In particular, I had made a bad initial assumption: that everybody who can create content of any type can write heterogeneous files to removable media. It's a simple assumption, but it has its roots in a clear understanding of the basic premise of the Windows or Mac hierarchical file system. Oddly enough, only a few of the participants could do this, which made saving significant portions of their digital possessions (without ongoing help) unlikely.

Perhaps the biggest surprise that came out of the first study was a profound sense of digital benign neglect, as symptomized by cycles of accumulation and loss. The majority of the study participants were struggling with computer viruses, half-installed software, and an impenetrable forest of digital content (some their own, some downloaded, some shared with others), along with partially implemented strategies for taking care of it. It wasn't that the participants in the first study didn't want to keep their stuff; it was more that they had neither the time nor expertise to do so. Using eerily similar metaphors, when asked about digital loss—either real or hypothetical—many of them said things like, "*I mean, if we would've had a fire, you just move on.*" (Marshall, Bly, and Brun-Cottan, 2006) and (Marshall, 2011)

It appeared that the long term disposition of participants' digital belongings would be left to fate.

Study 2. Personal data begins to lead a life of its own

The second study that I think of as a bellwether of larger trends in personal digital archiving took place around 2007, and was done in collaboration with Michael Nelson from Old Dominion University and his then-grad student, Frank McCown (Marshall, McCown, and Nelson, 2007). They were interested in reconstructing web pages and web sites using what they called Web Infrastructure, a collection of caches and archives that span many web resources (Nelson, et al., 2007; McCown, Nelson, and Van de Sompel, 2009). So what we were looking at together was online loss; we were interviewing and surveying people who had lost stuff they had stored on the Web.

The way this loss came about was what was surprising. The study participants had lost their online content largely by losing track of it, by not understanding the terms and conditions of the services they used, by storing it in proximity to illegally stored stuff, but not by what we generally think of as the most immediate path to loss: hardware failure.

The British Library conducted a study with similar findings: "*nearly 70% of [reported data loss in the home] manifested itself in an inability to find information; by comparison … [only] 8 % [of loss was] due to hard drive failure*." (John, et al., 2010) But if you ask people about backing up this online content, they will invariably tell you that there's no need to do that: everything online that matters has gone through local storage. Photos are taken off the digital camera and stored on the hard drive; video is produced locally, using local tools, as are documents. And these local copies are usually considered the digital originals: they are the highest fidelity and the most complete.

In fact, something else was happening when we first saw this result: a certain amount of circular reasoning was inherent to this personal information management strategy. Yes, the local copies were the highest resolution, but a certain amount of curation had taken place when this material was moved online—a service (and implied audience) was chosen; items were selected; metadata was added; and collections were organized. Furthermore, the online content was subject to ongoing stewardship as the items began to have lives of their own in the context of the larger digital resource. For example, photos in Flickr were selected by others to be part of thematic collections; comments were added; views were accumulated that reflected a picture's popularity. Meanwhile, the local originals languished, their actual location forgotten.

We concluded that study participants had lost important online assets through a variation on the theme of benign neglect.

Study 3. Ownership and control of online assets is messy business

Frank Shipman and I began the third series of studies in early 2010. We began the study series because we had a suspicion that ownership and control of online user-contributed material was different than it had been in a heterogeneous print/digital era, which would have an effect on both personal archives and institutional archives of personal content.

For example, I remember using stills from the movie *True Stories* in a talk I was giving in 1994. It took considerable effort to appropriate these images: I found a book about the movie; I scanned the pages containing the images I wanted to use; I did some primitive manipulation to get the pictures I wanted; I created a floppy disk with the image files on it; and I shuttled the media to over to a local business so they could print the images as 35 mm slides. Once they came back (several days later), I put them in my slide carousel. It was sufficiently painful that only a few of the images I used in talks were not my own to begin with. Today I do the same thing by doing a quick image search and a copy-paste into a PowerPoint deck. So does everyone else (as I have verified in recent interviews). Our personal archives have overlapping elements as well as casually appropriated and repurposed content.

Frank and I were convinced when we started that implicit social norms around saving, reusing, and archiving other peoples' content are coalescing. Surely people were curating content that wasn't their own, and they were expecting others, both individuals and institutions, to be doing so as well. But what surprised us was how a broad range of study participants were able to articulate their own rules and expected social norms.

For example, in data we gathered in mid-2010 about photo reuse (Marshall and Shipman, 2011), we saw participant reactions like this: *"[Reusing photos from the Web] is okay most of the time. The only time I would think it isn't would be when the main focus is of someone you don't know. Like when people email out the People of Walmart photos. Those are taken by people who don't know the person in the photo and posted."* This heuristic was far more specific, thought-out, and interesting than we had anticipated.

The remainder of the chapter focuses on these results, and what they mean when they are taken together with the results of the earlier studies.

Why the emphasis on reuse? Why would it matter that content we harbor in personal archives originated from online sources? We tend to think of personal archives as having crisp boundaries, and that we own everything we keep. But because people are such facile users of digital material (and keepers of appropriated stuff), the boundaries can be blurry. Some stuff is intentionally gathered to be part of an archive (e.g. photos of oneself, one's friends, one's family, or of important events); these photos may have been taken by someone else. Other stuff becomes an accidental part of a personal archive (e.g. a funny video clip that's circulating or a movie still appropriated to use in a talk).

Of course we can identify purely technical problems associated with reuse; for example, single-instance storage may be more efficient than storing copies. But that's not really why we're looking at reuse and the social norms that surround it. Rather, it helps to think about personal archives as something *live*, as resources, as material we (and others) might draw on in both the near- and long-term. In other words, personal archives (and archives in general) will become sources for future creative efforts. Reuse is an important motivation for maintaining a personal archive.

It is also likely that at least some of our personal online content will be part of an institutional archive (such as the Library of Congress's Twitter archive). Once such archives have been assembled, they will become a broader resource, and material in the archive is likely to be reused. Who will reuse the material and how it will be reused is part of an on-going social negotiation. Certainly reuse is at the heart of the uproar around the Library of Congress's Twitter effort—if preservation were the only consideration and no access was planned, objections would be harder to justify.

Of course, use has always been a factor in the construction of archives. But digital archives are fundamentally different than their physical predecessors. Not only is access far less constrained and rarified, but also the use of digital materials can be as extensive and imaginative as is permitted. In other words, reuse of materials—possibly re-reuse—will go far beyond historical research or social science analysis. This marks a distinct change from a time when you'd have to physically go to the archive, use a finding aid to establish what you wanted, and access the personal materials through the keeper of the archive.

If reuse is commonplace, and as legal scholar Larry Lessig (2008) suggests, that it is the basis for modern creative efforts, our personal archives—and, more importantly—institutional archives of personal material will be replete with 'lightly owned' content. An understanding of social norms will help inform both policy and design.

Emerging social norms

In his book *Code*, *Version 2.0* (2006), Lessig noted that peoples' behavior is governed through interaction between market forces, law, constraints imposed by technology (as well as affordances offered by technology), and social norms. While there's no denying that the first three factors have a pronounced effect, it is clear that the fourth, social norms, dominates most people's sensibilities when they decide whether or not they should save or reuse a photo turned up by a search engine.

"Will I get caught?" "Will the photographer mind?" "Are the subjects in a compromising position?" "Do I really need this photo?" All of these questions seem to float through the user's mind; in an instant—the amount of time it takes to copy-paste the image into a PowerPoint file—the doubts have been dispelled, and possibly forgotten. They only resurface if a take-down email appears, and even then, only seem to have an effect if legal action is imminent. "Will I get caught?" Probably not: most audiences are small, and most photographers only detect reuse if photo credit is already given. If the photo becomes part of the user's personal archive—and any potential reuse is deferred—no second thought is usually given.

As of May, 2010, 70% of the world's digital content was user-contributed; needless to say, that proportion has probably risen since. In other words, reuse is not limited to formally published content. Furthermore, reuse itself should be broadly construed: as Lessig has pointed out, much reuse may fall under the rubric of *remix*, a practice which results in derivative works that substantively change the intent and context of the appropriated material. Thus content creators are reusing less formal efforts and more narrowly-shared material, and they are using this material in ways the original owner is not likely to have imagined. And even if content creators seek permission before reusing user-contributed material, they are likely to engage in additional reuse without ever asking again.

Unfortunately, although Creative Commons labeling is a great system, designed for exactly this set of circumstances (Boyle, 2008), it is used and understood less frequently than it might be. For example, some participants in the study I'm about to describe think Creative Commons simply means that the labeled content is 'in the commons' and may be reused without further thought.

Thus for most purposes, social norms are emerging through successive cycles of use, reuse, modification, repurposing, and take-down notices. And although I'm about to tell you that many people in our studies adopt the aspirational attitude that permission should be granted prior to reuse, if we examine actual instances of reuse, the same participants would rather ask for forgiveness if they are caught rather than going through the potentially time-consuming and frustrating process of asking for permission beforehand.

A series of studies of ownership and control

Almost three years ago, Frank Shipman and I decided to look into these social norms by performing a series of studies. In particular, we wanted to characterize what these social norms were, discover when they break down, find out whether and how they vary across media types, and reveal what peoples' concerns were when archival institutions absorbed personal content from online sources. By doing this, we hoped to better understand peoples' attitudes and behaviors about content ownership and control; we also wanted to find out what people thought about archives as a source for repurposed content. So far we have looked at 6 media types: microblog posts (specifically Twitter tweets) (Marshall and Shipman, 2011a); personal photos (Marshall and Shipman, 2011b); book reviews; educational videos; comedy podcasts; and recorded videoconferences.

The other studies I have discussed took an ethnographic approach to answering their central research questions: we conducted lengthy interviews with a relatively small number of participants in an effort to reveal what they actually did with their personal digital belongings; we engaged with them over their own computers, storage media, and files. In this case, we wanted to reach out more broadly, to recruit a greater number of people (even if it meant we couldn't delve as deeply into what individuals did), to investigate attitudes as well as practices, and to compare answers across media types and specific user actions in a more structured way. Our publications about the studies describe the approach in greater detail.

To explore this notion of digital ownership, we used a technique familiar to legal scholars: hypotheticals that systematically vary a situation's fact pattern (Rissland and Ashley, 1986). The technique relies on a description of the basic case or scenario, followed by a series of 'what ifs'. In our application of the technique, participants rated the hypothetical propositions using a 7 point Likert scale (i.e. a 1 rating means *disagree strongly*; a 7 rating means *agree strongly*) as a single aspect of the fact pattern changed.

Let's say the central scenario posits a candid photo of a Halloween party-goer dressed up like Wonder Woman; a photographer who attended the Halloween party posts this photo to his Flickr account and tags it "Wonder Woman". This tag enables a comic book fan who is looking for pictures of Wonder Woman to find the photo. Now we can vary aspects of the fan's subsequent actions to reveal the edges of what the participant finds acceptable. For example, the first hypothetical might say, after the comic book fan downloads the photo, he stores it on his local disk in his Superheroes folder. The next variation might say he stores the photo in his public directory on Dropbox. Two subsequent hypotheticals might explore the difference between the fan reusing the photo by posting it to Facebook and reusing it to illustrate his blog. Varying facts this way enabled us to explore the ethical edges of the situation and to see where participants agreed and disagreed as well as investigating what they thought was acceptable behavior. We focused the questions on common actions such as storing, sharing, publishing, and removing different types of content.

Besides using the scenarios and hypotheticals, we also asked the participants to report on their own behavior. Of course self-reported behavior is tricky; we applied various tactics to keep participants from answering in a strictly aspirational way (by describing what they would do rather than what they have actually done). We kept these questions simple and neutral. At the end of each study we also asked one or two questions about belief: for example, is it okay to reuse photos that you find online? Why or why not?

The participants in these studies were young (in their 20s and 30s), but younger and older people were also part of the study population. About 1/3 of the population were students, and over 90% had gone to college (about 60% had college degrees). Thus the study population was young, well-educated, and probably under-employed. Many were freelance workers or office workers who spent much of their day in front of a computer screen. *These are the people who are likely to be in the vanguard of content creation and reuse—exactly who we wanted to reach in these studies*.

Although they were social media-savvy, participants were also remarkably diverse. Some claimed to spend most of their time reading and watching; others participated more actively, by blogging, tweeting, facebooking, and gaming. What we discovered was that although almost all of the participants reported that they shared and published online content, they had a broad definition of content in their minds. Profiles for social media, dating, and shopping, for example, all counted to participants as published content.

When we asked about sharing content, the participants exhibited nonchalance about reusing other people's material, and this was particularly true if something was funny or informative: "*[I share] music, interesting or funny pictures I come across, videos, jokes.*" Others passed on material they regarded as helpful, including coupons and good deals: "*[I share] information about …deals I find, or interesting articles I come across.*" Some even admitted to passing on explicitly copyrighted material. "*[I share] movies and videos. [I share] pornography and video games*." As we might expect, participants feel relatively comfortable saving anything they encounter online, regardless of the media type involved and their relationship to the content. Anyone can save any user-contributed content they find. But do they?

Here, practice is divorced from attitude. While downloading photos belonging to others is relatively common (72% of participants say they do so at least sometimes), saving reviews locally is far less so; participants rarely even save their own reviews. These join the other online ephemera, useful and used, but they are not saved nor extensively curated. In fact, the only non-positive reaction we elicited from participants was when we proposed hypotheticals that limited the ability to save content locally. For example, allowing a review's author to save her own review, but not the comments that had accrued on it, elicited a largely negative reaction. Similarly, limiting a tweet's author to save only his side of a Twitter conversation garnered a far more negative reaction than other hypothetical situations.

With this in mind, we can look at reuse, which is bound to be controversial and to have a more nuanced sense of where the boundaries are. Figure 5 shows four responses to a question about the last time a participant remembers reusing a photo.

boundary	*quote*

Of me	*“Someone tagged a picture of me on facebook and i saved the picture and put it into my album because it was a nice picture*.”
Of friends and family	*“my friend took a picture of my son and her son. i reposted so everyone can view from my family*.”
Of an event I plan to attend	*“Earlier this week I downloaded a picture of a dog wearing a party hat for a story I was doing on my pet blog about an event coming up. It was a photo included in a press release by the store holding the event*.”
Of a place I’ve been to	*“I couldn't find a photo I'd taken on a trip, which I wanted to use in a Facebook album, so I found a photo of the same landmark on someone's blog and republished it in my album*.”

Figure 5. The last incidence of reuse the participant recalls

What we see is a slippery slope of acceptability. In agreement with what we saw in the hypotheticals, social distance matters: even though I didn’t take the photo, *it’s a picture of me*. Or it’s a picture of my son. Or it’s an event I’ve attended. Or it’s an event that I plan to attend. Or—and here’s where distance begins to increase, perhaps to a breaking point—*it’s a place I’ve been*, and it’s a picture of a landmark, not a person.

People navigate a complex moral and ethical terrain when they reuse pictures. They may apply what they know of the law—or what they think they know of the law—and come up with a notion of the public Web as a place that is conceptually in the public domain. There has been ample discussion of limiting risk by simply not posting content; participants echo this discourse by saying things like, “*if you don't want your picture or face to possibly end up being seen by future bosses, kids, friends, parents whatever, then don't post it.”* In this regimen, Creative Commons is literally that, a commons, a place where content is offered for reuse without restriction, rather than a labeling system.

Although participants occasionally take their cues from the technology, using ‘share with’ buttons as a signal that it’s okay to reuse the content, more commonly they try to reason from the creation context or the photo’s intended genre as a means to intuit the implied permission to reuse. As we saw earlier, there is more sensitivity around personal photos than there is around landscapes, particularly photos that are not intended as artistic statements, but rather are simple evocations of vacation memories. If you’ve gone to Paphos and I’ve gone to Paphos, and you’ve taken a picture at the restaurant with the live pelican, and I eat there too but forget my camera in my hotel room, there would seem to be no reason why I can’t use your photo in Facebook to illustrate my travel story. It’s probably just as out-of-focus as mine would be, and with my small audience of friends, it’s unlikely I would be caught.

The closest participants come to echoing a Fair Use proposition is when they reason from reuse context. They intuitively know that commercial reuse is more controversial than non-commercial reuse, that derivative works may be a legitimate type of reuse, and that the particular motivations of the re-user matter. In practice, it seems that reuse may be mediated in a large part by need—when the participant actually needs a photo (instead of being handed a hypothetical situation), he or she is apt to rationalize reuse. Furthermore, the content’s history—has it been around much?—plays into the equation.

Most interesting are factors that arise from participants’ experience of life online. Will the reuse alter subsequent perceptions of the content’s veracity or importance? After all, if it is reused, it will appear in more places, and may seem truer or more significant because of its ubiquity. Is the reuse mean or malicious? There seems to be a fine line between ‘funny’ and ‘making fun of.’ Several participants cited PeopleOfWalmart.com as a website that crosses that boundary. Is the reuse misuse in the sense that it perpetrates fraud? When participants think about this type of reuse, they are more apt to think of a Photoshop job in which they’re made to look 20 pounds heavier than they are of phishing, or other more blatantly illegal deceptions.

A case for institutional archiving

When I argue that most people seem to approach the curation of their digital belongings with a mixture of benign neglect and unrealized plans to do better, I am in no way maligning them. We all are interested in keeping at least some of our digital stuff for the foreseeable future. In fact, my colleagues at the Library of Congress report that people have shown considerable interest in the Library’s outreach efforts and classes (see Chapter xx),and the Internet Archive has customers for its new personal archiving service. I regularly hear of new Internet startups with either long-term backup or archiving as part of their core mission; they have all perceived a need and detected an interest.

Why don’t I believe that, with this instruction and guidance, people can and will do it themselves? There are three reasons I’m skeptical:

1. *The disaggregation of necessary skills*. To be good personal archivists (even amateur ones), people need to perform curatorial duties (creating viable metadata, assessing long-term value). They need to perform the regular and extraordinary IT tasks that come with personal information management, everything from normal maintenance like installing software to heroic rescues like recovering from malware or device driver failures. On occasion, they may need to be media type experts too: is it better to store personal photos in RAW format or JPEG? To make matters more complicated, the family member who knows how to use Photoshop may not be the same family member who has an abiding interest in personal archiving, and neither of them may be the family member who has the IT expertise to perform everyday system maintenance. Add in the occasional duplication of roles and consultations with outside experts, and the result is an extensive roster of stakeholders and potential conflicts between them.

2. *Trends in personal data storage*. The storage picture is increasingly complex, but complex in a good way. We have more capable devices (our phones are all we need some days) and more special-purpose devices (think of Apple’s ecosystem of pads, pods, minis, and desktop-filling devices). There are more places in the cloud to store things and more things to store. Content that was once transient now may be kept indefinitely (for example, we can store everything from Skype conversations to screen captures of our latest raid in World of Warcraft), and content that was once personal is now shared via a growing array of social media services (at one time, Flickr was the dominant venue for photo sharing; now Instagram is the place-of-the-moment). There are more ways to publish the fruits of any creative effort that might cross our minds (some communities like DeviantArt have been going strong for over a decade; YouTube is a place where one can waste whole afternoons or seek emergency instruction for peeling a pomegranate; Pinterest changes the focus from creation to curation). The ecology of on- and

offline stores is complicated and unstable (Odom et al., 2012). I have interviewed people who pay for services they no longer actively use, simply because (like a physical storage locker) they don't know what to do with what they've stored there.

3. *The overwhelming tendency toward benign neglect*. Digital possessions accumulate rapidly. To properly curate them, we would probably need to drop everything we were doing and just devote ourselves to our legacy. At this point, it is important to recall the difference between *personal archiving* and *archiving personal material*: in the first instance, the owner of the material must set aside time and marshal other resources (the technical and curatorial skills I refer to above) to address digital belongings that may be of uncertain personal value; in the second instance, an archivist's considerable skills and an institution's resources can be brought to bear on the task and the material's value is seen from a broader cultural perspective. The poet's email acquired by the British Library that Donald Hawkins discusses in Chapter 10 is a perfect example of this distinction. Although over the last few decades, there has been considerable hand-wringing in some quarters about the disappearance of the literary letter (Arnold, 1999; Geoghegan, 2010), efforts by institutions like the British Library to preserve the email of important literary figures belie this fear. Yet for an individual, this task (preserving his or her own email) is onerous; email is often stored in a way that is opaque and figuring out a way to keep it safe is not offset by its overall personal value. The alternative (finding and saving specific high-value messages) is also time-consuming and usually seen as not worth the trouble. So usually personal email is simply left in place and its owner just hopes for the best: benign neglect at work. And certainly, in the end, for the individual archiving his or her own stuff, creation is just more rewarding than stewardship.

These three trends suggest that public institutions have a role to play in archiving personal digital belongings: if they don't intervene, there is no way to predict what will survive and what won't. But recent efforts, for example, the Library of Congress's initiative to archive the public Twitter feed, have been met with skepticism and even a certain amount of resistance. The skepticism stems from the feed's perceived lack of long term value. To quote a study participant, "*Things like facebook, twitter, google+ [should not be archived]. Some contain personal data that has no historical or archival significance.*" The resistance arises from privacy advocates, for certainly institutional archiving of social media violates an expectation of *privacy through obscurity*.[2] On the other hand, this initiative is consistent with Twitter's terms and conditions, and we must ask ourselves, "How much control do we have over data we have contributed to a site that stores it for free?" If we examine people's concerns, it's not that the Library of Congress is storing social media. Instead it's the potential for access and reuse. But it is that potential—the value of yesterday's social media for research, for art or everyday use—that may be the best justification for the cost and invasiveness of institutional archiving.

[2] According to Fred Stutzman, "*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?*" (Stutzman, 2010)

Frank Shipman and I have been interested in probing these attitudes as part of the series of studies I described earlier. In the six studies to-date, we always included a media-specific set of three hypotheticals based on this scenario: "*The Library of Congress is acquiring the public portion of the Twitter feed | Flickr | Amazon's book reviews | YouTube | iTunes educational recordings | popular iTunes podcasts, dating back to the site's origins. They are planning to provide access to the archive*." (We used the Library of Congress as a proxy for any public institution that would undertake a large-scale archiving effort because we felt that participants were likely to be familiar with the Library of Congress as an institution.) Then we posed three standard hypotheticals about how the archive might be accessed: by researchers now, by everyone now, or by everyone in 50 years. The results of these cross-media comparisons are described in Marshall and Shipman (2012). We also probed concepts like anonymity, satire, attribution, before-recording consent, creator permission, and removal from archives, many of which are analogous to those in models of Fair Use. (Sag, 2012) We wanted to see how sensitive our participants were to immediate public access (which is not that far from the initial situation: presumably the same works are available to the public through the services in which they were originally shared), and what kinds of limitations would ease any concerns they had.

If we look at the data for the immediate access scenario, participant opinion falls into two clusters: Tweets, photos, and reviews form the first cluster; participants seem less sanguine about immediate access to these collections. For tweets, in particular, the distribution is notably bimodal: some participants are very much opposed to immediate access, while others are mildly favorable. Educational recordings, podcasts, and YouTube videos form the second cluster; these are far less controversial. Most participants feel that immediate access would be acceptable. The reason for this seems straightforward, especially in the light of open-ended responses that are included in some of the questionnaires: the more personal the medium, the less acceptable public access is.

What would happen if we limited the access to the collection to researchers? For simplicity's sake, we did not define researchers, preferring to let participants use their intuitions. This limitation apparently eases risk in the participants' eyes, since most of the negative reactions have been attenuated. Tweets and photos have the most negative responses; educational recordings, the most positive.

Finally, we looked at the element of time—what would happen if we deferred access for 50 years? Twitter and reviews summon the least enthusiasm, perhaps because they are textual and most readily subject to retrieval and analysis. Time seemed to assuage some of the privacy concerns for visual media.

Open-ended questions about what should be excluded from this type of institutionally-curated collection revealed that participants assumed four perspectives that were similar to those we found for reuse by individuals. Figure 6 shows these four perspectives.

Creator	**Content/media**
• *permission* • *credit* • *privacy*	• *inherent value* • *veracity* • *harm/bias/malice*

Technology	**Legal/social**
• *authorization* • *settings*	• *public record* • *public domain* • *social good*

Figure 6. Four participant perspectives on institutional archiving

Unsurprisingly, it seems that participants were more concerned with potential effects of public access to these hypothetical collections. The prevailing concerns from the creator's perspective were permission and privacy. If participants took the content's perspective, they seemed to consider the content's value, weighed against the countervailing costs (supposing the content was, say, untrue or harmful). I was happy to see that a considerable number of responses brought up the overarching social good of such a collection. Only a few participants assumed the position that there were already technology-based mechanisms in place to handle these problems (e.g. Facebook's privacy settings). Figure 7 shows examples of the participants' responses to an open-ended question about institutional archiving of social media content.

Permission/Privacy *"...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"* *"Facebook pages, Twitter accounts, and any other social media ... It is that person's private account and should not be messed with."* *"…Videos are a lot more personal than anything written on paper so they should be treated more cautiously."*	**Value/Veracity/Harm** *"...No one cares what status updates someone in Colorado writes about the sandwich they ate for lunch."* *"… 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."* *"I think social media that contains racial bias or any kind of prejudiced based content shouldn't be archived."*
Existing mechanisms *"profiles indicate the user's express wish not to have their information accessible to everyone."*	**Law/Social Good** *"Once something is on the Web, it belongs to the Web users."* *"We are creating culture and history. No matter how some people feel about a certain subject or genre, nothing should be excluded."*

Figure 7. Examples that illustrate dominant concepts of each of the four perspectives

Although we suspected that ownership boundaries are fuzzy and that social norms are evolving, we also saw that content type and genre mattered. How personal the content is perceived to be has a substantial effect on how sensitive participants are to institutional acquisition. Furthermore, participants' familiarity with the media type in question, and how much experience they have with it in their everyday lives seems to make a profound difference in how they react to the hypotheticals we pose. Most importantly, we saw that collection building and access need to be teased apart; people aren't making this distinction on their own.

It is clear that public institutions have an important role to play in archiving social media. Between personal benign neglect and the onerous nature of the curatorial effort, it is unlikely that individuals will be able to do it themselves. Participants in our studies readily acknowledged role of public institutions in this immense stewardship effort; one participant said: "*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."* It is also evident that considerable effort still awaits us as archival institutions set the public's expectations about what should be in these collections; who should be able to use them and for what; and whether creators should expect long term plans to maintain their identity and to enforce attribution. As the Preserving Virtual Worlds project reminds us, even conservative approaches to obtaining permission have an associated cost. (In that effort, they tried to obtain in-world permission to archive content, and their overtures were met with hostility. In the end, they only achieved a 10% permission rate (McDonough et al., 2010), one that is heartbreaking for those of us who have set out on these missions with the best of intentions.)

In the end, it is important for us to acknowledge and accommodate to the enormous changes to personal archiving wrought by the emergence of digital content and modern personal information management practices. Physical metaphors (think of cardboard boxes in the attic or the office file cabinets) have broken down as digital content is stored in many forms, in an increasing number of venues, under the control of many stakeholders. Just as digital content has evolved, so too have digital practices. The ease and primacy of sharing, publication, and reuse has changed the nature and scope of personal archives.

Over the course of the studies I describe in this chapter, I have undergone a sea-change myself. I started out naively believing that addressing 'the personal archiving problem' was simply a matter of understanding individual practices and developing a system or a service for people to use to archive their own files. Over time, I have come to appreciate the complexity of the problem: personal digital archiving will require more than well-designed technology to become a force that is more powerful than benign neglect. Policy, education, and public and private efforts are necessary to realize the inherent sociality and reach of today's personal archives.

References

[1] (Anderson, 2004) Anderson, C. The Long Tail. *Wired*, Issue 12.10 - October 2004.

[2] (Arnold, 1999) Arnold, M. Making Books; Pen in Hand? Maybe No More, *New York Times*, January 21, 1999.

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

[4] (Geoghegan, 2010) Geoghegan, P. Epistles at dawn: the dying art of letter writing. *UK Guardian*, June 23, 2010.

[5] (James, 1890) James,W. 1890, *The Principles of Psychology*, New York: Holt, pp. 402-458.

[6] (John et al., 2010) John, J. L., Rowlands, I., Williams, P. and Dean, K. Digital Lives. Personal Digital Archives for the 21st Century >> An Initial Synthesis (Digital Lives research paper, March 3, 2010, Beta Version 0.2).

[7] (Lanham, 1993) Lanham, Richard A. *The Electronic Word: Democracy, Technology and the Arts*. Chicago: University of Chicago Press, 1993.

[8] (Lessig, 2006) Lessig, L. *Code: And Other Laws of Cyberspace, Version 2.0*, Basic Books, 2006.

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

[10] (Levy, 2001) Levy, D. M. (2001). *Scrolling forward: making sense of documents in the digital age*, Arcade Publishing, New York.

[11] (Marshall, 2011) Marshall, C.C. Challenges and Opportunities for Personal Digital Archiving. in (Cal Lee, Ed.) *I, Digital: Personal Collections in the Digital Era*. Chicago, IL: Society of American Archivists, 2011, pp. 90-114.

[12] (Marshall, Bly, and Brun-Cottan, 2006) Marshall, C. C., Bly, S. and Brun-Cottan, F. 2006. The Long Term Fate of Our Personal Digital Belongings: toward a shared model for personal archives. In *Proc. Archiving* 2006, 25-30.

[13] (Marshall, McCown, and Nelson, 2007) Marshall, C. C., McCown, F. and Nelson, M. L. 2007. Evaluating personal archiving strategies for internet-based information. In *Proc. Archiving 2007*, 151–56.

[14] (Marshall and Shipman, 2011a) Marshall, C.C. and Shipman, F.M. Social Media Ownership: Using Twitter as a Window onto Current Attitudes and Beliefs. *Proceedings of CHI'11*, Vancouver, BC, May 7-12, 2011.

[15] (Marshall and Shipman, 2011b) Marshall, C.C. and Shipman, F.M. 2011. The ownership and reuse of visual media. *Proceedings of JCDL 2011*, New York: ACM Press, pp. 157-166.

[16] (Marshall and Shipman, 2012) Marshall, C.C. and Shipman, F.M. 2012. On the Institutional Archiving of Social Media. *Proceedings of JCDL 2012,* New York: ACM Press, pp. 157-166.

[17] (McCown, Nelson, and Van de Sompel, 2009) McCown, F., Nelson, M.L., and Van de Sompel, H. 2009. Everyone is a curator: human-assisted preservation for ore aggregations. In *Proc. DigCCurr 2009*.

[18] (McDonough, et al., 2010) 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.

[19] (Nelson, et al., 2007) Nelson, M.L., McCown, F., Smith, J.A. and Klein, M. 2007. Using the Web infrastructure to preserve Web pages, *International Journal on Digital Libraries, Special Issue on Digital Preservation*.

[20] (Odom, et al., 2012) Odom, W., Sellen, A., Harper, R. and Thereska, E. 2012. Lost in translation: understanding the possession of digital things in the Cloud. In *Proc. CHI 2012*, 781-790.

[21] (Rissland and Ashley, 1986) Rissland, E.L. and Ashley, K. Hypotheticals as Heuristic Device. *Proc. Strategic Computing Natural Language*, Marina del Rey, California, May 1-2, 1986, 165-178.

[22] (Sag, 2012) Sag, M. Predicting Fair Use. Ohio State Law Journal, 73 (1), 2012, 47-91.

[23] (Sellen, 2001) Sellen, A. & Harper, R. (2001). *The myth of the paperless office*, MIT Press, Cambridge, MA.

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

[25] (Surowiecki, 2004) Surowiecki, J. 2004. *The Wisdom of Crowds*. Little, Brown.

[26] (Thorngate, 1987) Thorngate, W. (1987). On paying attention, in *Recent trends in theoretical psychology*, eds Baker et al., Springer-Verlag, New York, pp. 247–263.