Challenges and Opportunities for Personal Digital Archiving

Catherine C. Marshall

Until recently, the question of what people saved and valued centered on physical artifacts (e.g., print photographs, correspondence, and family records). It is only in the last decade or so, with the rise of inexpensive digital recording equipment (e.g., cameras), and equally inexpensive storage, that people began to accumulate significant collections of digital belongings. Using qualitative data gathered across a range of different personal digital archiving studies, this chapter examines the kinds of digital belongings people keep, how they think about deletion and loss, and how these notions translate into strategies and tactics for keeping digital material safe, both at home, and in the cloud. The chapter also explores elements of digital stewardship and personal collection building practiced by everyday people with an eye toward guiding them to a set of best practices to supplement what they already do.

It is sometimes difficult to recall that, as recently as the mid-1990s, personal digital media was still an exotic concept. In an article that appeared in the *New Yorker* in the fall of 1995, journalist John Seabrook took pains to explain to the magazine's well-educated readership what was implied by the alien concept of a home page. He wrote:

> In the simplest terms, [a home page] is . . . a place on the Net *where people can find you*. . . . Although building home pages or Web sites . . . is *mainly a commercial enterprise*, it doesn't have to be. It's also a way

> to meet people. . . . You can link your home page to the home pages of friends or family, or to your employer's Web site, or to any other site you like, creating a kind of neighborhood for yourself. And you can furnish it with *anything that can be digitized*—your ideas, your voice, your causes, pictures of your scars or your pets or your ancestors.[1]

From today's perspective, this explanation seems self-evident, almost quaint. Yet it is important to reflect on why it seems so, what dates this explanation. That people would amass such enormous quantities of valuable personal digital information was at that time inconceivable; computers and the data they were used to create, process, and consume were mainly the province of the business world.

Furthermore, there were few online venues for storing and sharing personal material—you either had a home page or you did not. The countless social media services we have grown to rely on—for example, photo sharing services such as Flickr, video sharing services such as YouTube, or blogging services such as WordPress—were yet to be offered. Person-to-person sharing was possible via email; but formatted email and email with attachments were still rarities in the consumer world. People were neither accustomed to having a persistent presence and online identity, nor were they used to storing their digital belongings anywhere online except for occasional accounts they purchased from providers such as AOL or Compuserve.

It is interesting to probe Seabrook's assumption about the origins of the material that people were sharing on their home pages. In the past ten or so years, many consumers have quietly migrated from using print and analog media to creating and viewing their own digital content; it is easy to forget that when John Seabrook wrote this article, digital materials were produced by a time-consuming and not wholly satisfying process of digitizing physical artifacts. There were few means of readily producing common kinds of emotionally enduring artifacts like digital photos or movies, and those that were available were still relatively primitive and difficult to work with. Formats had not stabilized; digital image resolution was low; and personal digital media could consume a substantial proportion of a normal person's storage resources. Consider, for example, that in 1995, an inexpensive digital camera cost eight hundred dollars or so, produced grainy 640 by 480 pixel photos, and reproduced colors with unsatisfying fidelity. Furthermore, early digital cameras were bulky and impractical. Laptops were not ubiquitous, so there

was often no easy way of uploading photos when the photographer was traveling. To make matters worse, a 3.5-inch floppy disk, the most common removable storage medium, held only five or six TIFF-format photos.[2] Post-processing capabilities were similarly primitive: image correction algorithms were still in their infancy.[3] Thus, a state-of-the-art digital camera circa 1995 gave the buyer little to recommend it over a ten-dollar single-use film camera.

But perhaps the most profound shift is social rather than technical; even this recently, people primarily shared and treasured print photographs; archival "originals" used to save the photos and to produce additional print copies were stored as negatives. Thus we have only had a decade to become accustomed to photographs that are born digital, stored as files, and shared and managed electronically.

It is vital to remind ourselves of the magnitude of these changes as we go on to consider the state of digital archiving around that time.

In January 1995, Jeff Rothenberg published a groundbreaking article about digital archiving in the widely read magazine *Scientific American*. The article not only described the problem and proposed an early technical solution; it was also regarded as a prescient call to arms to the computer science community. We should start thinking about digital archiving now, the article seemed to say, before we begin a fall down the slippery slope into a digital dark ages.

A scenario describing a hypothetical situation circa 2045 formed the linchpin of Rothenberg's article. The concrete problem of digital archiving was framed like this:

> The year is 2045 and my grandchildren (as yet unborn) are exploring the attic of my house (as yet unbought). They find a letter dated 1995 and a CD-ROM. The letter claims that the disk contains a document that provides the key to obtaining my fortune (as yet unearned). My grandchildren are understandably excited, but they have never seen a CD before—except in old movies—and even if they can somehow find a suitable disk drive, how will they run the software necessary to interpret the information on the disk? How can they read my obsolete digital document?[4]

One document. Stored in an unreadable format on an obsolete storage medium evoked by old movies. And a contrived story about how it got there and why we would want to see the document again.

On one hand, this situation is breathtakingly dramatic (the author's legacy is all there, indecipherable, in a single document!); on the other hand, more pedestrian than what is true a little more than a decade into this imagined future. Instead, the average consumer is already overwhelmed by the sheer volume of his or her digital belongings. It is not just one document; it is thousands and thousands of pictures, hours and hours of undifferentiated digital video footage, music both purchased and shared, personal finances, the beginnings of what will one day be extensive medical records, email messages important and trivial, and countless other files representing day-to-day interactions with the computer and with other people by means of the computer.

Nor is it accurate to assume one central nexus of storage: today's computer user is unlikely to store his or her files in one place or on one medium; a personal collection is more apt to span many repositories, storage media, and file systems. Even now, most of us have lost track of online accounts and removable media. If a terabyte disk can fit into the palm of one's hand and costs less than a week's worth of groceries, it is easy to see how such a device can be readily misplaced or forgotten. Furthermore, because we have so many portable devices (music players, phones, USB keys, digital photo frames) and so many different audiences (our friends, our families, our coworkers and colleagues, our accountants and doctors, strangers with similar interests) it is easy to see how it is in our interest to distribute our personal collections far and wide.

Nor does the scenario represent the complexities of our informal personal information technology (IT) arrangements. Notice that the document's author is both its owner and the person who wrote it onto the CD in question. Access is unquestioned—the document is neither encrypted, nor password-protected, nor stored in an external account—and the CD has apparently either been stored in felicitous circumstances or has been refreshed in the intervening years. In any case, access to it is unimpeded either by intent or by the ravages of time.

In short, to represent the problem as embodied in a single document in a central store, curated invisibly (if minimally) over time misses many of the most important aspects of personal digital archiving.

Because the article emphasizes access to a single file, already to hand, Rothenberg's solution takes a principled computer science approach: to render a stored digital file at the highest possible fidelity, in a form true to how it was created, complete emulation of the document's original computational environment—the hardware platform, the operating system, and the application used to create the document, at the very least—is necessary.

Emulation is a sophisticated solution for keeping digital objects alive. Through emulation, current hardware and software may exactly imitate the behavior of legacy hardware and software; full emulation of the architectural layers beneath the application allows the application software (or an emulated version of the software) to run just as it did originally, and allows this software to interpret and render the digital file exactly as intended.

Emulation is a complicated archiving strategy because it requires so much information about the original hardware and software environment. Not only must hardware emulation duplicate the processor's behavior (for example, its speed and the operations it supports), which must be fully specified; it must also duplicate the behavior of any necessary peripheral devices and the drivers that access them. An operating system and other platform-level components such as compression/decompression software, fonts, and applications libraries must run on top of the emulated platform. Although emulation may seem straightforward, it is important to remember that computational environments are complex for many reasons and it is difficult to fully specify them; for example, a bug in the original operating system may be paradoxically essential for producing a desired behavior, yet it seems unlikely that the bug would be included in the formal specification. Although they seem straightforward, environmental elements such as fonts may include bits of code to specify, for example, how adjacent characters are kerned.

Given a desire to fully preserve the digital object, it is easy to see why emulation was such an appealing building block for digital preservation. And while emulation is clearly possible—the very nature of computation makes it so—it consumes resources, and may not be necessary. As David Levy

pointed out several years after the Rothenberg article appeared, ultimately preservation methods should be dictated by expected use.[5] Will the digital object be edited and continue to evolve? Does it have properties that mitigate against keeping it in a living format? Who will access it and what will they do with it?

So let us step back and look at Rothenberg's scenario from a modern perspective. Consider that as of May, 2010, the photo-sharing service Flickr hosted well over 4.6 billion photos; as of April, 2009, the photo-sharing service Photobucket hosted more than 7.2 billion; and the social networking site Facebook hosted more than 15 billion unique photos.[6] Shifting the focus from Web 2.0 services to people, by 2008, estimates place digital camera adoption in the U.S. at 67% of all households; individual households (those who have adopted broadband) store about 14 gigabytes of photos, a figure which is almost doubling yearly.[7] These photos may be centralized on a disk drive (apart from the rest of the user's digital belongings), or they may be scattered far and wide on web services such as Facebook, Flickr, or ImageShack. They may be backed up or otherwise duplicated, or they may be unique; they may have descriptive metadata (tags, titles, or meaningful file names), or they may be only accessible via the files' time stamps.

Multiply this problem by other common media types and digital artifacts, and by burgeoning web storage services and cheap removable personal storage devices like USB drives, and you can begin to see the scope of the problem. Issues such as copyright enforcement and technologies such as digital rights management software bring legal considerations to the table. Because some of the digital material represents what is most private to an individual (be it financial or medical records or emotionally charged love letters), it is reasonable to expect any discussion of digital archiving to take up matters of protecting archival stores against intrusion and to consider subtle issues of privacy, especially where families and friends come into the picture. It might be just as traumatic if *everything* survives the passage of time intact as it would be if *nothing* does.

In short, personal digital archiving is a complex problem that raises social, technical, and legal issues; it is unlikely that this problem will be solved by a single application, web service, service provider, institution, or policy, and certainly it is a problem that extends far beyond accurately rendering obsolete formats. There is a broad frontier of questions and approaches to

explore. While it would be most satisfying to offer a system architecture, a mechanism, a representation, or even a coherent story that addresses digital archiving with a holistic solution, it would be misleading to do so. So instead, I will frame the problem by introducing three overarching questions that have arisen from the changes that have occurred since we have grown into digital beings with significant things to archive:

- What does a typical personal digital collection contain?
- What is the technological basis for storing, maintaining, and preserving this collection?
- Which stewardship practices do people bring to bear on the problem and which best practices should they be taught to supplement existing practices?

With each of these questions, I will discuss particular elements of an adaptive solution—one that does not commit to a specific path of technology adoption, one that does not assume the growth of a specific media type (e.g., digital video), or presuppose drastic changes in human behavior.

Figure 1 adapts the digital library framework presented in *Going Digital: A Look at Assumptions Underlying Digital Libraries* to structure these three fundamental questions as arising from the interactions between digital stuff, technology, and human activity.[8] Given this framework, any personal digital archiving effort will necessarily involve a digital collection (or more accurately, an accumulation of digital belongings that neither has the distinct boundaries, nor the acquisition policies of a more formal collection); technology for storing and sustaining digital belongings; and digital stewardship practices.

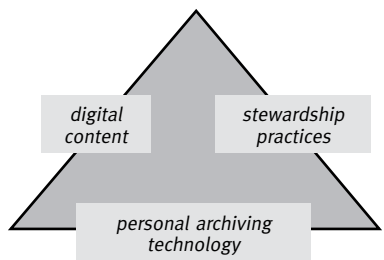

Figure 1. A personal digital archiving framework.

In other words, personal digital archiving will boil down to nothing more profound than deciding what we should keep, how and where we should store it, and what sorts of work people will have to do to keep their digital collections alive. At first blush, the answers seem as simple as the questions. As a strawman collection policy, we will keep everything. A secure central repository containing self-describing digital objects protected by sophisticated access policies will provide people with a facility very close to a digital safety deposit box that uses the emulative principles that Rothenberg—and later Henry Gladney and Raymond Lorie—had in mind.[9] Finally, personal digital stewardship can be modeled after the best practices we have developed to keep important individual collections alive when they are donated to archives, museums, and libraries.[10]

We can dust our hands off and go home, having assigned the question of best practices to professional archivists, the storage question to computer scientists, and having punted on the collection question entirely. Or can we? Let us look at each question in turn; then let us ask ourselves "Is that all there is?"

What Is in a Personal Digital Collection?

Certainly the most basic question about personal collections centers on what we should keep.

It is very tempting to throw up our hands and say, "Storage *is* cheap. Is there any reason not to keep it all?" Even if our personal digital belongings are accumulating at a truly breathtaking rate, sharp reductions in storage cost make it unlikely that we can break the storage bank. Although we have all had the experience of filling up our hard drives, we usually regard that occurrence as a signal of obsolescence, as a warning sign that it is time to buy a new computer, or at least install a new, more capacious, disk drive.

And why not keep everything? There are plenty of reasons to do so. First—and most importantly—it is very difficult to assess an item's or media type's future worth many years before the value is proven. Although we are confident that some digital assets are indeed worth keeping—the great American novel we have begun writing (and perhaps have returned to at different times through our adult years); our wedding photos; our children's graduation videos—in short, the documentation of life's milestones and

evidence of the sweat off our creative brows, there are others of more ambiguous worth. Yet often the candid snapshot turns out to be our favorite photo twenty years down the road. Journals from our high school years that are embarrassing when we are young adults may be rediscovered and valued when we are older. They may even have larger cultural value, or at least emotional significance, to future generations. Conversely, the insurance policy we put in a safe place to grab in case of fire is fodder for the shredder once the policy has expired. Bank records that were a necessity during an IRS audit might be a liability in another situation. Mindful of their value, we often carefully store such records, and do not rediscover them until they are worthless. Even items of great emotional worth may degrade: wedding photos may not be treasured forever if a marriage ends.

Although individual items are of varying value, keeping everything is more or less in line with how many people implicitly manage their personal computers. Rather than deliberately deleting files, they may either leave files behind when they transition to a new computer with the assumption that they can revive the old computer if they discover they need something, or they move the files wholesale to the new machine without examining them.

Three participants in an interview study who were all experienced computer users described this transition process in very similar language:

1. "Whenever I get a new machine, I just transfer everything over. And I just dump the old box."
2. "[When I buy a new computer] I transfer everything. . . . [The computer] is the same [except] it's faster. I should take the time to clean it up at that point, but [I don't]."
3. "I usually transfer stuff from PC to PC when I buy a new one . . . [I] transfer [files] directly by ftp [and put them] usually in my own folder structure." [11]

Let us think about this strategy: what it means is that in the abstract, your computer is a living archive; at any given moment, the computer that you are currently using could possibly store every digital item that has ever crossed your path. Of course, this is seldom true. External storage, web services, old hardware, computers used on the job rather than at home, and any number of forgotten elements of one's personal digital environment may hold a portion of one's personal digital belongings. But for the time being, let us

not be so literal. Let us say that a "keep everything" strategy is logistically possible and results in an indefinitely large store of all of your digital stuff for all time. If people are left to their own devices, this mode of accumulation seems to reflect the natural order of things.

It is also important to realize that deletion comes with an associated cost. Why fight against intellectual gravity? After all, deletion is difficult, thankless work; it is cognitively taxing to decide whether to keep something or toss it and the immediate return is minimal. And as yet, there is no Nobel Prize or Oscar awarded for maintaining a neat, well-pruned file system.

Again, in practice, deletion is a far less systematic process than one would hope for or even imagine. It is far more likely that when they are doing something else, people will notice an unfamiliar file name, or encounter an item that seems to be of no persistent value, or assess a piece of email as spam, irrelevant, or no longer interesting, and delete it on the spot (often without even looking at it). Symptomatic of this behavior, during interviews, participants will spontaneously delete a few files as a symbolic act, declaring their distaste for clutter. However, this unpremeditated act does not mean that a participant is committed to spending the next week sorting through her files. For example, in a 2005 study, as we watched a couple going through their hard drive, one of them said to the other: "I don't know what that is. You might as well delete it as far as I'm concerned." In other words, there is no need to look; it is unfamiliar, so it is unlikely to be valuable. In spite of the fact that there are probably many such low-value files stored on the computer, the participant somewhat arbitrarily deletes the one that is currently in view, frequently vowing to go after more such files soon.

During the same study, we were observing a participant go through a folder during an exercise in which we asked people to look for the oldest personal file stored on their hard drives. The participant in question encountered a candidate file, did not immediately recognize what it was, opened it, and remembered, thereby setting off a longer reflection about deletion and decision:

> "This could've been a seminar or something. Now I remember—[the file's contents describe] what was going to be in the seminar and I didn't go to it. Wow. Yeah. I haven't looked at this stuff in a long time. I mean, it would be sort of interesting to get the pop-up, make a decision on it, and then let it go . . . [in the future] I will become a lean, mean organizing machine."

Yet he did not delete this particular file once he remembered what it was, nor did he delete any of the others he assessed as clutter.

In another study—this time focused on researchers and their personal scholarly materials—one participant was going through his files as I watched. He found a directory on the server replete with files he decided were no longer valuable. As I continued to observe him, the researcher began to delete files without even looking at their contents, only the file names and dates. Sometimes he did not remember their purpose, and sometimes he did:

> "The server also stores tons of stuff from when I was working on [project x]. We stored all kinds of shared stuff out there so we could access it. And I need to go through and clean that up because there's gobs of junk out there that should just get deleted . . ."
>
> [I ask what a particular file contains.]
>
> "I don't remember. That's why it should probably all be deleted. . . . These I think are datasets from various runs. Yeah. This is—no, this is installation files. For various versions of [the software developed for project x]. All this should go. Yeah. These are installation files. They should all go away. They're useless."
>
> [He begins deleting files.]
>
> "None of these things mean anything. Except for this. But the line analyzer stuff. I'll never use it. So let's just get rid of that. This stuff should go away. This looks like an archive of the [final version of a conference paper]."[12]

Most experienced computer users realize that they delete items in a somewhat arbitrary manner. What varies is whether they plan to someday become more methodical in their curatorial efforts, or whether they accept the vagaries of their own stewardship practices—whether they plan to become lean, clean, organizing machines, or whether they see the endeavor as more or less hopeless.

During another study (Marshall et al., 2006), a participant acknowledged that his digital stewardship practices were more or less arbitrary. When I asked him whether he ever got rid of digital stuff, he said:

> "Yes, but not in any systematic manner. . . . It's more like, I have things littering the desktop and at some point it becomes unnavigable . . . . A bunch of [the files] would get tossed out. A bunch of them would get put in some semblance of order on the hard drive. And some of

> them would go to various miscellaneous nooks and corners, never to be seen again."

Another participant, when asked how his files were organized, confessed:

> "I keep telling myself that maybe one day I'll basically do the computer equivalent of spring cleaning. I'll just find all these scattered directories and files and sort of clean them, create a fresh hierarchy of 'here are my pictures,' 'here are my movies,' 'here are my documents,' 'here is my music' and get them all cleanly laid out along those lines. And I just never seem to find the time."[13]

There is another, more principled, reason to keep everything. This perspective stems from the idea that the computer will eventually serve as the ultimate memory prosthesis.[14] In other words, the computer's storage is equated with the capacity of human memory, and by extension, deleting corresponds to forgetting.[15] Thus, if we adopt the memory prosthesis perspective, deleting files means that useful context is eliminated, and we lose the ability for every digital item to act as an index to the other digital items that occurred around that time, in that place, or with those people. Deleting files is tantamount to deliberately making a hole in one's memories, inducing amnesia, and potentially reducing the prosthetic power of the computer. But the advantage of keeping everything comes at a price.

Recall the reasons we would keep everything:

- It is difficult to assess value in advance;
- Keeping everything aligns well with current practice;
- Deletion is itself a cognitively demanding exercise;
- People are seldom methodical about culling their files, so why even try; and
- A full chronological and contextual record is essential for using one's archives as a memory prosthesis.

Just because it is easier keep to everything than to cull it, does this mean that there is no virtue in the natural falling away of digital belongings through time's gradual erosion? As we examine what people do, a puzzling pattern emerges: people seem to be relying on disk crashes, technology failure, and periodic obsolescence as a way of pruning their collections. It is not that

loss does not bother them; it is rather that loss makes their collections more tractable. The accumulated weight of these digital belongings is swept away, so that they can focus their attention on the present. Three quotes from separate interviews conducted in the course of the study cited above confirm that this is a common perspective:

> "You want to know the truth? If I blasted my 11,230 emails away, I wouldn't be that bad off probably. Because I'd be able to work on new ones coming in."
>
> "If my hard drive was gone, it really wouldn't bother me all that much, because it's not something I need, need. I just thought it would be nice to keep in around in case I have another [school] assignment just like it."
>
> "I mean, there's plenty of stuff that I've lost that I thought used to matter to me, but not so much. I used to be a real America Online chat guy back in high school. And I'd chat with people from all over the world and one of the cool things is that your logs would get saved and I actually met one of my girlfriends online . . . I could literally go back a year and just look at old chats that we would have. . . . I switched PCs and I just kind of forgot to transfer those files over. Or maybe I wasn't able to. Maybe the formats were incompatible with the new version. I don't remember what happened, but at this point, I've clearly learned not to care. I mean, I thought I would've cared. It might've been nice if I still had them."[16]

This cycle of accumulation and accidental lose may underlie the thousand logistical explanations that consumers offer for failing to back up their computers. In the end, people may be unhappy about data loss, but they shrug it off, all too frequently saying exactly the same thing: "I mean, if we would've had a fire, you just move on." Of course, if you had as many fires as computer crashes, you would look for the arsonist among your friends and family members; you would not simply move on. But we can readily identify some countervailing reasons why we would not keep everything. First and foremost is that although storage is cheap, human attention is far less so. Furthermore, as we will see later in this chapter (and in other chapters of this book), stewardship is more than simply storing digital belongings once on reliable storage; stewardship requires continual attention to the items and media in a collection: Are the formats still active? Is the storage medium still good and still up-to-date? Has virus protection successfully eliminated

all threats? Keeping everything places an enormous tax on our stewardship resources. Finally, even if we are careful not to equate computer storage with human memory, keeping everything is impractical from the perspective of the current legal system; we are often instructed to discard certain items.

Thus, from the collection standpoint, what we are looking for varies with value, and that value is not a guaranteed attribute of an item. Most certainly—at the ends of the spectrum—we are looking for the ability to safeguard the things we really care about (even if we are sometimes wrong) and we are looking for the ability to permanently expunge the things we know we never want to see again (which, again, is different from the items of ambiguous worth). But even then, we are left with the vast bulk of the items in the center—those of uncertain value, those that may form the linchpin of our memories or may be not worth looking at in ten years. In other words, most challengingly, for those items in the middle, we are looking for the digital equivalent to benign neglect.

This stratified view of value suggests that a single personal collection may merit several different value-related strategies for keeping digital materials safe. There are at least four categories of value:

Known high value items. There are certainly items in every personal collection that are known to be of high value (whether value is assessed correctly or not) and thus warrant full archival treatment. These are the things we know we want to keep, and although they vary from person to person, in practical terms, these are usually small in number and identifiable. Most digital archiving strategies to date have been oriented to these items.

Medium value subcollections. Most people can express the value of types of things: I would not want to lose my photos or I would feel bad if I lost the videos I have taken of the kids. Yet examination of these subcollections reveals that they are large and not of uniform value. Some loss would be tolerated, and, over time, some of these items will turn out to be of high value. Certainly people will maintain these things in the present—that is, these items are used in everyday practice and people will copy them from computer to computer. Safety will generally be ensured by making ad hoc copies of the items.

Lower value subcollections or media types. These are subcollections, media types, and items that are of more ambiguous worth. For example, we have

seen in the quotes included earlier in the chapter that for many people, overstuffed inboxes or shared music falls into this category. The bulk of the items are a burden rather than a clear benefit, yet there may be items of real value hidden among the detritus. People often express their ambivalence to these items by not backing them up nor copying them to removable media; instead they tempt fate. These are the items left behind when a new computer is purchased. Yet, it seems not worth the effort to delete individual items, and there s enough of worth that one certainly would be reluctant to simply expunge the entire subcollection. Along with items of medium value described above, these subcollections may be preserved through use. The more the items are used, the more valuable they are likely to be, and the better their chance of survival. These items are good candidates for the heuristic approaches that I have sketched out elsewhere.[17]

Items of known liability. As many researchers have asserted, there are items that people would like to "forget"—that is, they would like to delete them from their digital holdings and be certain they will never reencounter them, regardless of the circumstances. Unlike archival deaccession, this deletion is not related to value, but rather is tied up in emotional or legal liability. Hence, deletion must ensure the item is not forensically recoverable.[18]

Storing Personal Digital Archives

At first blush, it is tempting to specify a single central repository for storing our personal digital archives. Then we have control over formats of individual items and can either migrate them or store them in a self-describing way, or provide emulation capabilities to decode them well into the future. Furthermore, sophisticated access policies would enable us to crisply say who has access to what. Our accountant can get at our financial records; our doctor can get at all of our medical records, and our insurance company can get at some of them; our families can look at most of our photos, but perhaps not forward them to people outside of the family. It is easy to come up with a repository that in principle satisfies all of our archival needs.

In practice, however, there are other forces at work. People have their own rationale for putting portions of their overall digital assets in different places. Data safety is an essential side-effect of this de facto distributed storage, but it is unlikely to work the other way around. That is, a centralized archive

is unlikely to offer all of the advantages people are already realizing from storing their data in specialized repositories. They will not have the audience they command with Flickr, YouTube, or a blog server; they will not have the functionality offered by various email providers. They will not have local control, as they would by putting the assets on a home server. And although they may have the same level of security they have for financial records stored at their brokerage, they will not be able to conduct transactions.

The follow excerpt from an interview (conducted via Skype's IM functionality) is telling:

> "[11:09:24 PM] g says: [There are] 6 [online places where I store things] in all. 1.) school website, 2.) blogspot, 3.) wordpress.com (free blog host, different from wordpress.org), 4.) flickr, 5.) zooomr(for pictures, they offer free "pro" accounts for bloggers, but even for non-pros, they don't limit you to showing your most recent 200 pics only unlike flickr), 6.) archive.org
>
> [11:10:42 PM] cm says: I ask just because you seem to have stuff in a lot of different places (so far two different blog sites, flickr, youtube, msnspaces, . . . maybe yahoo?) . . .
>
> [11:11:07 PM] g says: oh right . . . youtube because people always tell me that they don't feel like downloading my quicktime files from archive.org."

This type of strategy is typical: the participant has stored items in multiple places (sometimes the same items; sometimes different ones) with a plan in mind. She has well-articulated reasons for her choices (Zooomr offers her a free "pro" account, which does not limit her to displaying her two hundred most recent pictures like Flickr does; yet this does not keep her from storing photos on Flickr as well). She is conscious of the fact that different stores reach different audiences (elsewhere in the interview, she states that MSNSpaces reaches a Taiwanese audience that her other blogging accounts do not); this diversity is reflected in her choice of language on the different sites (English or Chinese or a mix of the two). Interestingly, she has already begun losing track of what is where. This confusion is important insofar as a personal archiving strategy that revolves around de facto distributed storage also demands that the user somehow keep track of the far-flung digital assets. Indeed, this challenge is underscored by the fact that data loss is more commonly tied now to losing track of where things are stored and

the policies and practices of various storage providers rather than to local crashes and catastrophes under one's own control.[19]

It is very tempting to ignore these trends and offer a centralized service to implement the functionality and services we associate with a capable digital repository that takes a long view of storage. But a centralized service will solve only part of the problem at best: it may well be an attractive place to store high-value items of a certain sort—photos, for example, or other types of personal media. But as I have pointed out, other high-value items will be stored in other types of repositories—repositories with capabilities well suited to the type of items stored there.

Hence, what is called for is federation at the metadata level, via a catalog or some other repository that records not only metadata, but also the health and characteristics of other repositories. A catalog can variably store a variety of types of surrogates for individual items, varying from the items themselves (in the event primary storage goes belly-up) to metadata that represents salient features of the remote items. This enables authoritative versions to be stored in the most appropriate place, while the archive keeps track of where the asset is, and any information needed to access it and ensure its ultimate health. This way, our actual distributed store can involve everything from items that are stored as attachments to free email (such as Hotmail, Gmail, or Yahoo mail) to medical records stored in specialized for-pay vault software. Implementing an archive this way allows policies and agreements to be tracked centrally without requiring sensitive information—or media directed at a particular audience—to be similarly centralized.

Network bandwidth use is minimized; security problems are reduced; access is parceled out very naturally; and the sort of loss that we are already observing is minimized. This approach begs the question of format, deferring decoding to access time. This is perhaps a dangerous strategy, and there is no in-principle reason that a user could not be warned of potential obsolescence, since metadata can be used to track storage format and any fonts and codecs necessary to render the item.

This method also acknowledges the need to care for items representing a range of values differently. Perhaps we would want a more fail-safe method of caring for high-value items and to merely ensure that several copies of lower-valued items exist. It also seems that this type of approach

can implicitly acknowledge the human tendency toward benign neglect, by gently allowing gradual loss of items of ambivalent value.

Archival Best Practices and Personal Stewardship

As well-known people begin to donate born-digital materials, the engines of their creation (e.g., laptops and desktop computers), and common storage media to libraries, museums, and archives, professional archivists are developing best practices for the stewardship of these personal collections.[20] But the question is, how relevant are these best practices to the consumer at home who has neither the resources, inclinations, skills, nor time to apply them? Can they be scaled back to fit the home user, or are they irrelevant and arcane?[21] Or, even if they are easily understood, are they pragmatically possible from a resource perspective (that is, does anyone have the time to put them into practice)?

What we see when we visit people's homes and workplaces confirms any suspicions we might have about this diverse group of users: they have the best of intentions—they do not actually want to lose data in an uncontrolled way—but they also have other things on their minds. For example, one study participant showed us a yellowed newspaper clipping on a stand near her computer. The year-old technology tips column, "Saving Files with a CD-RW Drive," gave the home user detailed instructions about how to write files to a CD. When asked if she had used the instructions, the woman allowed that no, she had not. But she intended to, when she had time. I feel fairly certain that she has not followed the instructions, unless a family member has taken over and done it for her.[22]

Because among the digitally prolific there is an implicit tendency to rely on periodic loss to keep uncontrolled growth in check, study participants are beginning to own up that they are not even sure they believe in digital stewardship. One technologically sophisticated study participant who had lost his personal and business websites because he had not run a recent backup admitted that he was not sure he wanted to take responsibility for the loss; furthermore, from his musings, it did not seem as though he intended to implement a backup procedure in the near future:

> "It's funny though. If you look at technology, it's just one of those things. I mean, whose fault is it? Is it the user's fault for not backing

> up? Or is it technology's fault for not being more tolerant and failsafe? In ten years, maybe hard drives and PCs will be so invincible and the Internet will be so pervasive that the concept of backing up will be quaint."[23]

When we talk about teaching archival best practices, it is not only important to identify the best practices or the mode of instruction; it is also vital to specify who will be learning them and who will be implementing them. There are several important things to remember. First, the home archivist is often not the same person as the home IT provider (although home archivist and home IT provider are both ad hoc, fluid roles). The woman with the "Saving Files" clipping next to her monitor is most certainly her family's archivist—at other points, she talks about saving photos, family recipes, and so on—but she is undisputedly not her family's informal IT support. That role belongs to her ex-husband:

> "I tried to install it [Firefox] and then John [her ex-husband] said, 'Don't install anything on your computer.' . . . I usually defer to John. Because he's the one that's got to come over and maintain it. So I have to make sure that it's okay with him. But Jack [her 18 year old son] . . . will just do whatever he wants."

From the opposite perspective, the capable IT person may have little interest in curating a family artifact. In other words, the person who is capable of creating the CD described in the previous example may not see the need to label it (as would be the impulse of the family archivist). In another interview during the same study, a college student held up an unlabeled CD at an angle so that she could see the contrasting textures that indicate what part of the CD has been written and told us, "It's kind of weird, but with some of these CDs, you can tell how much is written on it by looking."

It becomes apparent that in many home situations, archiving is a cooperative activity, one in which people take different roles with respect to the technology. People tend to rely on each other for informal backups of their own archival efforts. Canot identify someone in a photo? Maybe a friend or family member will be able to tell you.[24] Were not able to get a particular shot? Maybe someone else thought to bring a camera. Experience a technology failure? Perhaps the person you sent an attachment to still has it and can recover it for you. The digital stewardship we associate with personal archives is inherently social.

Again referring to the same set of interviews, a student who had lost an autobiographical assignment to a computer virus was able to recover her document from a relative:

> "Even my personal statement was saved onto that computer [the virus-infected laptop]. Then luckily, I also emailed it to my cousin, Camilla, at her house. . . . So I said, 'Camilla, do you still have my UCLA personal statement?' She's like, 'Yeah.' So I said, 'Okay, can you please email it.' So then that's how I actually got it back to this computer."[25]

It is not unusual for people to recover lost media socially, for example, from a recipient of an email attachment, or from a collaborator.

At this point, we must return to matters of scale. At the scale of the single digital artifact, individuals are rapidly improving at creating, recording, shaping, mashing up, sharing, and even saving one item at a time. But they are no better at keeping these things around as we scale up along various dimensions (as the individual items grow larger and more numerous; as the number of people we share with grows larger; as the number of devices we use increases; as the time of retaining items approaches decades). In other words, the same practices that allow us to handle one media file often cannot be applied to handle an entire collection.

As one frightening illustration, we posed the following scenario to some interview participants: suppose you had one hour before you knew your local hard drive was going to fail. What would you do? A surprising number started their panic reaction by saving one file at a time—for example, a participant might describe attaching an item that mattered to an email message (we were presuming mail services were still available). In practice, we know this one-at-a-time approach is not feasible.

Matters do not really improve as users become more sophisticated and are able to handle more files at once and move to the cloud. A participant in another study indignantly described losing her collection of personal journal podcasts after a service went under:

> "I hosted my podcasts early on on a free service called Rizzn.net . . . he then changed rizzn.net to something called blipmedia.com and then . . . he decided to sell blipmedia . . . and he never emailed people about it . . . suddenly the files were gone and the only news I heard about it was when I had to hunt online for what happened."

Thus, if we return to our earlier realization—it is easier to *keep* than to *cull*—we can further muse that it is easier to *lose* than *maintain*. And that, in a nutshell, encapsulates benign neglect as a personal digital archiving strategy.

Other strategies for keeping digital materials safe reflect current best practices. Of course, automating whatever practices we can routinize is the best way of taking advantage of the computational platform. There are already personal archiving services that have made substantial headway on providing individuals and small businesses with curation services and mechanisms.[26] It does not seem necessary to spell them out here, except to say that there is plenty of room to investigate services that take advantage of the so-called wisdom of crowds, since that is one point of leverage that is far more available in today's world. In essence, just as there are orders of magnitude increases in the number and genres of things we can create in a digital networked world, there are similar increases in our access to the fruits of other people's labor; if we are clever, we can harness the power of social networks to perform various communal organizing and labeling activities.[27]

Does a Collections-Storage-Practices Framework Cover the Territory?

Obviously, reducing the challenges of personal digital archiving to an interaction among collections, storage, and archival practices is too simplistic. Much falls outside of this framework, especially the question of use as time passes. How will people recover items from a very large distributed personal store, especially after they forget what they have and what they do not have?[28] It is important to note that recovery of digital assets from long-term storage is very different from implementing a generalized search mechanism (either adversarial or based on straightforward term frequency).[29] This is the point at which we will pay the price for our storage strategies.

The other notable gaps in this discussion stem from the rapid shifts underway in underlying technologies, shifts that may give us new points of leverage. For example, the replication mechanisms that are already the backbone of institutional stores may be brought to bear on the problem of how we synchronize content among our many personal devices.[30] Once this type of mechanism is in place, we may have a straightforward way of

scooping up copies of valuable digital assets as they are replicated and storing them in a repository that lies somewhere between a backup and an archive.[31]

But the main lesson we should take away from this discussion is not that a 'silver bullet' technology will come along and render the problem solved, or that a method of creating self-describing objects will allow our digital assets to be perpetuated into the indefinite future, or even that the first born-digital generation—today's kids—will have a much better understanding of appropriate stewardship practices, and hence will just know what to do. Instead, we should realize that there are many equally valid approaches to creating and maintaining personal digital archives and that all of our digital belongings do not require the same level of attention and protection.

An archive that is in essence a catalog will allow us to federate our digital belongings on the metadata level and will help us resolve issues associated with provenance. This kind of approach will allow different stores to evolve to meet different needs. After all, it is likely we will want to archive our medical records differently from how we archive our financial records. Furthermore, we may even want to devote different resources to our formal portraiture than we do to our cell phone snapshots (recognizing that we may be wrong about their relative value). Different methods thus may be brought to bear on preserving our medium- and low-value items.

Above all, any personal digital archiving solutions should acknowledge the human tendency toward benign neglect. For example, we might want to know that we are about to lose access to our financial records because we have changed banks; we may want our social networking profiles to be safely stored when a start-up's business model proves to be unworkable; we may want to be reminded of our encryption keys and of our many temporary stores.

It is perhaps disappointing that we are not tilting full speed at a digital memory box, a digital safety deposit box, or an infinite digital U-Store-It. The desire to centralize, to fully federate, to unify and standardize is understandable, but it also seems out-of-step with human nature. Although the world offers us plenty of scrapbookers and intergenerational storytellers, we have to remember that these phenomena are exceptional and exclusive: they lose the incidental, the candid, and the accidental—a significant portion

of the stuff of history that has to date been preserved through benign neglect.

Benign neglect is thus a means of transcending the vagaries of accumulation and incorrect assessments of value. Distributed storage reduces the vulnerabilities of centralization, and allows functionality and design to more fully reflect the genres of the stored media. Acknowledging that people will not at once become capable stewards of their own digital belongings gives us a realistic idea of what to expect. Benign neglect and intrinsic distribution can become instrumental in securing a digital future in which we neither keep everything, nor lose everything, nor become shackled by the need to sustain our growing accumulations of digital belongings.

Notes

[1] John Seabrook, "Home on the Net," *New Yorker*, October 16, 1995, 70 (italics mine).

[2] For example, the author still has twenty-nine photos she took on an Apple Quicktake digital camera on a 1995 trip to Graceland, Elvis Presley's home in Memphis. Most of these photos were stored as 225 Kb TIFF-format files, except for a few 900 Kb high resolution shots.

[3] Without post-processing, it is hard to discern the subject of many of the Graceland photographs; modern correction algorithms are capable of improving the photos to the point where they are viewable. This noticeable improvement in correction capabilities makes a compelling argument for keeping the original rather than a corrected version.

[4] Jeff Rothenberg, "Ensuring the Longevity of Digital Documents," *Scientific American* (January 1995): 42.

[5] David M. Levy, "Heroic Measures: Reflections on the Possibility and Purpose of Digital Preservation," in *Proceedings of the Third ACM Conference on Digital Libraries*, ed. Ian Witten, Rob Akscyn, and Frank M. Shipman III (New York: ACM Press, 1998), 152–61,

[6] http://techcrunch.com/2009/04/07/who-has-the-most-photos-of-them-all-hint-it-is-not-facebook/ (accessed May 29, 2010).

[7] From the report *Home Servers and Consumer Storage*, prepared by Parks Associates, a market research firm (Dallas, Texas: July 2008).

[8] David M. Levy and Catherine C. Marshall, "Going Digital: A Look at Assumptions Underlying Digital Libraries," *Communications of the ACM* 38, no. 4 (April 1995), 77–84.

[9] Henry M. Gladney, "Principles for Digital Preservation," *Communications of the ACM* 49, no. 2, (2006) 111–16; Raymond A. Lorie, "A Methodology and System for Preserving Digital Data," in *Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '02), Portland, Oregon, July 14–18, 2002 (New York: ACM Press, 2002), 312–19.

[10] Emory Libraries, *Preserving Salman Rushdie's Digital Life* (2008), http://www.youtube.com/user/emorylibraries#grid/user/8A1D63F362925EA9 (accessed August 26, 2010).

[11] C. C. Marshall, F. McCown, and M. L. Nelson, "Evaluating Personal Archiving Strategies for Internet-based Information," in *Proceedings of Archiving 2007*, Arlington, Virginia, May 21–24, 2007 (Springfield, VA: Society for Imaging Science and Technology, 2007), 151–56.

[12] Catherine C. Marshall, "From Writing and Analysis to the Repository: Taking the Scholars' Perspective on Scholarly Archiving," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '08), Pittsburgh, Pennsylvania, June 16–20, 2008 (New York: ACM Press, 2008), 251–60.

[13] C. C. Marshall, S. Bly, and F. Brun-Cottan, "The Long Term Fate of Our Personal Digital Belongings: Toward a Service Model for Personal Archives," in *Proceedings of Archiving 2006*, Ottawa, Canada, May 23–26, 2006 (Springfield, VA: Society for Imaging Science and Technology, 2006), 25–30.

[14] Gordon Bell and Jim Gemmell, *Total Recall: How the E-Memory Revolution Will Change Everything* (New York: Dutton, 2009).

[15] Evgeny Morozov, "Speak, Memory: Can Digital Storage Remember for You?" *Boston Review* (May/June 2010), available at http://bostonreview.net/BR35.3/morozov.php.

[16] Marshall et al., "Long Term Fate of Our Personal Digital Belongings."

[17] Catherine C. Marshall, "Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, and Institutions," *D-Lib Magazine* 14, no. 3/4 (2008), http://dx.doi.org/10.1045/march2008-marshall-pt2.

[18] S. Garfinkel, and D. Cox, "Finding and Archiving the Internet Footprint" (paper presented at the First Digital Lives Research Conference, London, England, February 9–11, 2009).

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

[20] P. Cohen, "Fending Off Digital Decay, Bit by Bit," *New York Times*, March 15, 2010.

[21] Certainly many people in this field have made sincere efforts to make archival best practices accessible and have thought about how individuals might implement them at home. See, for example, the Library of Congress's suggestions at http://www.digitalpreservation.gov/you/index.html (accessed May 29, 2010).

[22] Marshall et al., "Long Term Fate of Our Personal Digital Belongings."

[23] The participant was part of the study described in Marshall et al., "Evaluating Personal Archiving Strategies."

[24] Indeed, more recent social media services such as Facebook rely on just this sort of collaborative photo-labeling activity.

[25] Marshall et al., "Long Term Fate of Our Personal Digital Belongings."

[26] S. Strodl, F. Motlik, K. Stadler, A. Rauber , *Personal and SOHO Archiving (*New York: ACM Press, 2008).

[27] I feel compelled to emphasize the "if we are clever" portion of this assertion; it's easy to fall into the trap of thinking that crowdsourcing will solve everything. Obviously, it won't, and there are both good and bad examples of how it works (or to point out resources such as Wikipedia that are examples of failings as well as strengths of this approach). See S. Sen, S. K. Lam, A. M. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, F. M. Harper, J. Riedl, "tagging, communities, vocabulary, evolution," in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work* (CSCW '06) (New York: ACM Press, 2006), 181–90.

[28] And this does not even consider the question of the digital belongings that are passed from generation to generation, creating a gulf of meaning that becomes even more difficult to surmount. It is hard enough to curate a personal collection for one's own use and the use of one's immediate family and close friends, let alone for use by different generations who are unlikely to share enough context to reconstruct the meaning of particular artifacts.

[29] For a discussion of the challenges of such recovery, see Catherine C. Marshall, "Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field," *D-Lib Magazine* 14, no. 3–4 (2008), http://dx.doi.org/10.1045/march2008-marshall-pt1. See also E. Cutrell, S. Dumais, and J.

Teevan, "Searching to Eliminate Personal Information Management," *Communications of the ACM* 49, 1 (January 2006): 58–64.

[30] V. Ramasubramanian, T. L. Rodeheffer, D. Terry, M. Walraed-Sullivan, T. Wobber, C. C. Marshall, A. Vahdat, "Cimbiosys: A Platform for Content-based Partial Replication," In *Proceedings of the 6th USENIX Symposium on Networked Systems Design and Implementation* (NSDI '09) (Berkeley, CA: USENIX Association, 2009).

[31] See, for example, offerings such as Sharpcast (http://www.sharpcast.com) or Microsoft's LiveMesh (http://www.microsoft.com/livemesh).