David Gartner

Designing for Digital Archives

Elizabeth Churchill
Yahoo! Research| churchill@acm.org

with **Jeff Ubois**
Fujitsu Labs of America | jeff@ubois.com

Have you amassed a collection of photos and other media without quite knowing how to manage it? Have you spent hours trying to locate a precious or extremely important file? Have you ever wished you'd backed up your files after a computer crash?

More and more of our work and personal content is digital. And mobile, digital technologies like camera phones are changing the nature of capture and collection—what and how we collect. We are living in a world of continuous accumulation.

This is relatively new. Ten years ago fewer people had home computers, fewer services existed, and we weren't surrounded by all those appealing, shiny devices that promise to record our every action in case we want to take a step down memory lane or revisit an article written a while back to snaffle some useful content. Back then terms like "moblogging", "lifelogging," "microblogging," and "lifestreaming" were not in common parlance.

Ironically, this ease of capture and replication actually makes it more likely that we'll lose stuff. The sheer volume of data we are able to collect makes organization daunting and specific content difficult to locate. Frankly, the logically extreme vision of life as constant accumulation offered by Gordon Bell and his collaborator Jim Gemmell, with their MyLifeBits project, is apt to make anyone with old-time curatorial sensibilities erupt in hives.

Amplifying the challenge is the fact that content tends to accumulate in various places—on internal or external flash and other portable drives; on recording devices themselves (cameras, audio recorders, phones); and hosted at ISPs and by services like YouTube and Flickr. Few people have a centralized repository of all their stuff. We curate, consolidate, and/or back up randomly or not at all, and have muddled mental models regarding file formats, backup, and archive practices and services. Prospective retrospective—that is, imagining now what we will want to remember in the future—is hard; we have a limited ability to gauge such future value. So we have a propensity to defer decisions about whether something is worth keeping or not.

Consequently, most of us are what Microsoft's Cathy Marshall and her collaborators have called "lazy preservationists," who rely on "opportunism, optimism, and benign neglect." And most of us are living in a world of digital bloat, our untamed and insecure data strewn all over the place. We skip along on a wing and a prayer, explaining away catastrophes and rethinking data importance in the face of loss: "I guess it must not have been important if I lost it." Sometimes this kind of loss and revision is therapeutic. Sometimes it is not. Sometimes we spend hours reconstructing content or creating passable replacements. For our own archives this is personally troubling, but as a culture it is positively terrifying that our data and our memories are at risk.

Some see this problem as a commercial opportunity. GYMA (Google, Yahoo, Microsoft, AOL) are exploring the business of archiving, backup, and storage, and services; others, like Seagate's Mirra Personal Server, Apple's .Mac account, EMC's Mozy promise storage and a "data cloud" where our stuff will be safe … forever. Or until we fail to pay the subscription fee. Or until they have business or technical problems. Or, as happened to one of our own *interactions* columnists, some malicious miscreant masquerades as you and in a click of a button or two, deletes all your precious material. Under most terms of service agreements, users have no recourse and companies have no obligation to restore the "lost" material even if back-ups exist.

We need to develop a finer

appreciation for the risks to our data posed by "solutions" to other problems (such as DRM), and understand that data preservation is becoming a struggle with active adversaries—malware authors, political partisans, and scammers conducting phishing attacks. Commercial organizations have a mixed record as long-term custodians of personal artifacts and of cultural works.

So in the light of all this, what are some approaches designers and other stakeholders may be interested in exploring? After all, service, application, and interface designers will be the ones implementing the experience now, and thus have a direct impact on the future of our personal and collective digital memories. And who are the stakeholders whom we need to be talking to and designing with, for, and around?

Here are our top five clusters of points and questions on this emerging area. These are overlapping, and there are more, so consider these a seed list.

1. Guide users between backups, archives, and collections. Good design for archival services can help users make decisions based on anticipated future uses and perceived risks.

For starters, it is helpful to distinguish between archiving and backup. Apple's Time Machine, which is part of Mac OS X Leopard, is an interesting step in the right direction. People report learning that *a backup is not the same as an archive* when old (but important) versions of files have been overwritten by backup software whose check boxes were clicked (or not). The options the check-boxes offered required knowing the distinction. Perhaps systems need to ask questions like the following: "Are you sure you want to overwrite this file with all future versions?" Yes, that means *overwrite* it. Not store another version and keep track of all that you have done with the file.

Users must choose between a wide range of file format and compression options (think of ZIP, TAR, JPEG, MPEG, PDF...). Some are proprietary, some may be unsupported in the future, and some are "lossy," meaning file sizes shrink by reducing resolution. Purists in the archival community rule out the use of lossy compression (MP3 or MPEG 2) altogether when there are non-lossy options available (FLAC or JPEG2000). But for personal collections of audio and video, lossy algorithms may be the best way to limit storage costs. Systems that allow users to preview the difference, or that explain the implications of loss, may help.

As professional librarians and archivists know, you cannot have archives without curation. At a more personal level, psychologists view strategic forgetting as what constructing a (more or less) stable sense of self is all about. In this case, a question posed to the user might be, "Are you sure you want your kids to see this when they go through your archives?"

The importance of forgetting should not be lost on us. However, we need to guide users through these concepts with intelligently designed systems and interfaces if people are not going to inadvertently lose the digital materials they *want* to keep. Unfortunately, the consequences of bad decisions may be felt only days, months, years, and decades later. It is hard to learn best practices when there is this lag, so once again designers need to surface the results of choices and knock-on effects at the time of action.

2. Be involved in conversations about the differences between algorithmic search and human memory. Over time we may be able to follow Google's directive, *search don't sort,* because improvements in search algorithms and applications will eliminate the need to file content manually. This search-don't-sort perspective is also reflected in David Weinberger's book, *Everything is Miscellaneous*, in which he explains how the ordering of our collections can be reworked on the fly, as the situation demands. This argument is most compelling if metadata is well designed and standardized. So, for this approach to work, we should be active in communities where forms and standardization of metadata are discussed. Simply asserting that people can be less careful about providing metadata because search is improving is an unacceptably risky approach for materials that are worth saving.

A complementary approach is to leverage our understanding of the way in which human memory works—by recreating context to facilitate retrieval. This would entail providing time frames punctuated by memorable events (salient or regular events), congruent activities ("I was working on the Rosebud project when I took that picture"), and so on. The point is, what we remember is sometimes not the *searchable* content. In these instances we narrow the search space through circumstance reconstruction—a kind of semantic way-finding to the content… "*something from 2004 when Mum came to visit, so it must have been August and it was a picture and it would have been….*" Again, Apple's Time Machine in Mac OS X Leopard explores this, giving you a snapshot in time of your files. This is an appealing idea.

A lot of human information interaction is serendipitous, based on vague, ill-formulated, semantic associations not clear on text and numbers, and enacted as browsing, encountering, and being reminded—not explicitly remembering. A text-search string still does not find a figurative image, and file metadata are volatile. But reconstructing context is a powerful memory-jogger bringing back the abstract textual that goes with the recognized visual.

Search will also need to return results that cut across different media. Google's Universal Search, which provides results from video, images, new, local, and book search, is a step in this direction. Yahoo!'s OneSearch does this nicely for cell phones. Ask.com does it too, but prettier.

The world is waiting for the designer who can (re)create and implement the memory palaces and mnemonic techniques used by renaissance scholars and described by Frances Yates in *The Art of Memory*.

3. Data is dynamic, not static. The great promise of an archive is to assure long-term access to information. That sounds like stasis, but it isn't. To be effective over decades, archival systems need to migrate data from disk to disk, and in some cases, emulate the environments of the applications that use the data.

In considering personal data storage, we need to consider the easy migration of personal data from one location to another. But personal and social data are always evolving; they are not stable. Formats change, data migrates between storage methods and places, and security and access methods evolve. Smart organizations are looking to support users in their understanding of the consequences of that volatility. Services are beginning to take on the responsibility of educating users as well as funding research into data migration and fighting against format obsolescence (often by supporting current as well as legacy formats).

Digital rights management schemes that allow limited access today may fail in ways that allow no access tomorrow.

For designers these considerations may lead to uncomfortable practices. Refusing to innovate in favor of traditional practices and technologies; sticking close to the file system rather than adding a layer on top; and avoiding the unique in favor of the conventional as a way to support future users and avoid evolutionary dead ends all go against the desire to improve on past practice.

4. From personal to social data. Archives sit at the boundary between public and private data. Data that was once private may, through an archive, gradually be made public. That presents new opportunities and challenges the digital environment.

One opportunity is in cataloging, which is expensive for both institutions and individuals. When the individual is overwhelmed with too much content to name, tag, sort, and store, we could always harness the crowd, get the group to tag and organize. Crowdsourcing and services like Amazon's Mechanical Turk harness human intelligence to solve problems that computers find hard—like tagging and organizing and storing. Archiving is a collaborative practice, and it is going to become ever more so.

But this solution brings up another issue we need to keep in mind: Who becomes responsible for the content created through a collaborative enterprise, and how are ownership and responsibility for that content conceived of by the service providers? An article in Wikipedia is distinct from the contributors who created it, but if a photo that has been collectively tagged in a photo-sharing site like Flickr "belongs" to an individual who subsequently leaves Flickr, what happens to the content? Many people are crushed when the comments they have made on blogs disappear because the blog "owner" stopped maintaining the blog.

Relying on social approaches to archiving may be a practical necessity, but open archives must be built to withstand and respond to a wide variety of attacks, not only from individual malware authors, but from political partisans, abusers of copyright law, and even governments that wish to control access to historical records.

The Society of American

Archivists Code of Ethics states "archivists protect the privacy rights of donors and individuals or groups who are the subject of records." We need to think also about the "rights" and caretaking of the collectively created data. There are questions about ownership of the augmented data that need to be addressed. We need to create a place for discussion of practices around data augmentation with socially contributed metadata.

5. *Designing for sustainability.* We have heard much in the press recently about establishing provenance, considerations of authenticity and integrity, and content rights. Recent efforts from groups such as the Organization for Transformative Works address the trials of remix and fandom with their statement: "We envision a future in which all fannish works are recognized as legal and transformative and are accepted as a legitimate creative activity," wanting to protect fans, the work, the commentary, the history, and thus identity, "providing the broadest possible access to fannish activity for all fans." Access is certainly part of it, but as a secondary point preservation must be central; if the content is not maintained, issues of ownership and control are moot. Who wants to be in control of nothing?

Services and technologies bring with them responsibility if they are to be sustainable. Alfred de Grazia, a pioneer in personal digital archiving, has reframed the problem as one of "managing intellectual estates." The beneficiaries are not just the individual user, but also our culture as a whole, and our descendants. Part of the solution is in an economic model that can be used to sustain and encourage preservation and allow intellectual estates to be maintained. De Grazia focused on the needs of the academic arena. However, with many of us now producing portfolios of mixed-media content for work and being archivists of our own past and those of others, these points are clearly generalizable and more relevant to a broader audience today. As blogger Dave Winer put it, "With all possible humility, I'd like to tell you that a few days after I die my entire Web presence will likely disappear…And when my sites disappear, so will my uncle's. He died in 2003. His site is still accessible because I keep it that way." He points out that his uncle's thoughts may not be something the world at large cares about, but if Dave's uncle were a Nobel Laureate, it would likely change things. In the same post he also points out that most universities do not have a plan for archiving the Web-based content of their professors. Clearly, some folks need to be reminded that the Web is an extensible publishing platform, not an Etch A Sketch.

Digital technology makes it possible to extend the walls of the archive beyond a single space or person, as well as ensure preservation and access in locations around the world in what the Library of Congress is calling a "content stewardship network." Libraries, museums, and archives will need to collaborate with business interests to build lasting social structures that are sustainable over time. There is much work to be done and many stakeholders to be engaged and heard in the merging of content from multiple sources.

A Final Note

To close, it is worth pointing to Terry Kuny's 1997 paper that circled library science networks, warning of a coming *digital dark age* when our data will be lost and/or irretrievable unless we individually and collectively recognize the vulnerability of digital data and design better tools, procedures, services and policies. We say: Let's appeal to greed, fear, utopianism, and good design and make sure we prove him wrong.

ABOUT THE AUTHORS
Dr. Elizabeth Churchill is a principal research scientist at Yahoo! Research leading research in social media. Originally a psychologist by training, for the past 15 years she has studied and designed technologies for effective social connection. At Yahoo, her work focuses on how Internet applications and services are woven into everyday lives. Obsessed with memory and sentiment, in her spare time Elizabeth researches how people manage their digital and physical archives. Elizabeth rates herself a packrat, her greatest joy is an attic stuffed with memorabilia.

Jeff Ubois is exploring new approaches to personal archiving for Fujitsu Labs of America in Sunnyvale, California, and to video archiving for Intelligent Television and Thirteen/WNET in New York. He has been published in *First Monday*, *Release 1.0*, *Computerworld*, the *Journal of Digital Information*, and *D-Lib*, and he blogs at http://www.archival.tv.

DOI 10.1145/1340961.1340964