‘It’s like a fire. You just have to move on’:

Rethinking personal digital archiving

Cathy Marshall
Microsoft Research Silicon Valley

FAST 2008
27 February 2008
In Silicon Valley, the web was in evidence everywhere

sign for San Francisco dive shop circa 1995

from Avocadoh’s photo stream on Flickr
early web site

"Home of Friendly People & Polish Sausage"

Bremond is located in the Brazos Valley in the heart of Central Texas within 45 miles of Bryan/College Station, Temple and Waco. Bremond is also within 140 miles of Austin, Dallas/Ft. Worth and Houston.

Many immigrants from Poland came to Bremond in the 19th Century and their heritage can be seen in the many events planned each year. These include:

POLSKI DZIEN (Polish Festival)
ST. MARY’S CHURCH FALL FESTIVAL
DOWNTOWN CHRISTMAS FESTIVAL
CITY WIDE GARAGE SALE
FARMERS MARKET BEING PLANNED

Bremond boasts an active Chamber of Commerce as can be seen by the list of prosperous businesses in the city and nearby.

The community supports several churches and schools.
early homepage

Jim Bob, Bryn and the Boys

Jim Bob, Bryn, Zac and Ty

Sorry, there's not enough time to be cool!
Apple QuickTake digital camera
my trip to Graceland

29 mostly awful photos in tiff format...
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 uneamed). 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?

Jeff Rothenberg, “Ensuring the Longevity of Digital Documents” SCIAM, Jan ‘95
...his solution: emulation

“If I include all necessary system and application software on the disk, along with a complete and easily decoded specification of the hardware environment required to run it, they should be able to generate an emulator that will display my document by running its original software.”

Figure 9: Using emulation to read an obsolete digital document
fast forward to 2008
there are more than 2.2 billion personal photos on Flickr

and if that’s not enough, Facebook has at least 5 billion more...
It’s becoming obvious that our digital stuff is important to us

Premise: the writer offers $1000 for personal items, including strangers’ laptops. He gets wallets, pocket contents, wedding rings, but not laptops.

“At a Starbucks on Michigan Avenue, I approached a kid hunched over an ancient-looking laptop covered in skateboarding stickers. He thought it over and shrugged. ‘No way,’ he said. ‘I am this shit. Everything in here.’ A woman at the same shop said she hated hers. ‘But come on,’ she said. ‘Sell you my laptop? That would be like selling you my knees.’”

And how are we actually losing this digital content? (hint: it’s not format yet)
A skeptical reviewer’s comment

“Seriously: what’s the hangup? As long as I take out the photos and look at them every decade or so, it’s a piece of cake. We buy a new computer every few years, spend a few minutes moving our documents folder to the new machine, we’re done. You aren’t suggesting that, come 2054, nobody will remember how JPEG works?”
Translation: “why don’t we just do what our parents did—put the stuff somewhere safe and forget about it”

it worked for the cardboard box under the bed...
this ‘doing nothing’ is sometimes referred to as benign neglect...

which is more or less the fine art of just leaving well enough alone
“...neglect can sometimes be an artifact’s best friend.”

- G. Thomas Tanselle

“Statement on the Significance of Primary Records”
benign neglect would’ve worked better here

reel-to-reel tape used to archive rare vinyl records...

rare vinyl records
So, perhaps the solution that’s the most equivalent to the box under the bed is to shove everything into a big database now and decode it later…

“Bookcase now, in the ground later. Size is whatever you need.”

…but can personal archiving really be reduced to storage and self-describing digital objects?
emphasis: artistic intention

digital arts archiving

emphasis: emotional worth

archiving personal digital belongings

emphasis: coverage

internet archiving

emphasis: authenticity

records archiving

scientific data archiving

emphasis: scientific context

archiving institutional/library holdings

emphasis: part of the discipline
How can we find out what personal digital archiving is really all about?

by looking at what’s going on around us...
This talk draws on real data from real people and their stuff

- consumer field study in 3 cities
  what people save, where they keep it, and is it working?
- survey and interviews of people recovering lost websites
  the difference between network storage and local storage
- field study of researchers and their scholarly output
  the difference between researchers at work and consumers at home
- case study of a long-term email correspondence
  the difference between 10 years and 25 years
The first thing we noticed was how resigned some people are about losing their stuff.

They even wax philosophical about it.
“If [my email messages] were totally lost it wouldn’t be the end of the world. I guess that I don’t consider anything tangible, like, so important as an emotion or an experience, I guess I’m kinda of like a Buddhist.”

“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 it around.”

“I mean, if we would’ve had a fire, you just move on.”
On the other hand, some people aren’t that sanguine about losing all their stuff...
“If I lost my Gmail account and all my associated email, I’d probably have a schizophrenic episode or something. Because I use it for more than email. I email myself just important little chunks of data... [online email] makes it convenient for throwing files up in a sort of protected way."

“if Yahoo ever disappears then I’m screwed.”
saving files with a CD-RW drive

Q: My Windows XP computer came with a CD-RW drive but not a floppy drive. Is there a way to make a CD that can be copied to a floppy? Do I need special software?

A: Not all CD-RW drives are compatible with floppy disks. You’ll need to check the user manual for your CD-RW drive. If it’s compatible, you may be able to create a CD that can be copied to a floppy. Otherwise, you’ll need to use software that can write to a floppy.

Q: I recently switched Internet service providers from America Online to Comcast. With AOL, I could use the spell-checker in Outlook Express. Is there any way to activate it?

A: If you’re no longer using Outlook Express from America Online, you’ll need to download a spell-checker. One option is Microsoft Office Word, which includes a spell-checker. If you don’t have Microsoft Office, you may need to download a spell-checker from a free software source. Some spell-checkers are available online for free, such as the built-in spell-checker in Word or the WordPerfect spell-checker. You may also be able to download a spell-checker for your specific word processing program.
how do consumers believe they archive their digital stuff?

- they believe their backups are archival
- they move files wholesale onto latest PC
- they write files to removable media
- they use email + attachments
- they put files on media sharing sites
- they save old platforms
and sometimes they think someone else is doing it for them

“No, I’m not backing up our files—I’m just assuming that the F.B.I. is making copies.”
All of these methods have some things in common...

The people I’ve interviewed all assume:

- no further curation is necessary
- they can keep track of everything
- they can recognize the good stuff
- they’ll be able to retrieve what they want when they want it

but most of all

- they’re going to remember what they have!
personal digital archiving:  
4 challenges & themes
A skeptical reviewer’s comment

“Seriously: what’s the hangup? As long as I take out the photos and look at them every decade or so, it’s a piece of cake. We buy a new computer every few years, spend a few minutes moving our documents folder to the new machine, we’re done. You aren’t suggesting that, come 2054, nobody will remember how JPEG works?”
challenge 1: accumulation, asset value, and provenance

People have a rough time predicting future value. Digital stuff simply accumulates or is ruthlessly eliminated.

When asked when he ever got rid of digital stuff, one consumer participant 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 them 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.”
value is where principles and practices collide...

Folk wisdom...

- Copy stuff to keep it safe.
- Stay organized and keep clutter to a minimum.
- Back up stuff to minimize unintentional loss.
- Anything you get from the Web can be easily replaced.
principles & practices: make copies

principle: Copy stuff to keep it safe

[from consumer interviews] “I could burn it on CD but that’s – I’d have to look for a blank CD somewhere.” (theory v. practice)

[from lost website interviews] “I mean, the photos go off of my camera onto my computer before they go up to Flickr. So I always have master copies on my PC.” (which is the ‘original’?)

[from researcher interviews] “I'm very paranoid about losing data. So in addition to being on three computers, it's being backed up from two of them.” (is five enough? is ten too many?)
principles & practices: stay organized

principle: Stay organized and keep clutter to a minimum.

[from consumer interviews] A couple going through their hard drive while we watch: “I don’t know what that is. You might as well delete it as far as I’m concerned.”

[from researcher interviews] “there’s gobs of junk out there that should just get deleted... [e.g.] we've got log files from various test runs.”

[from consumer interviews] “[In the future] I will become a lean, mean organizing machine.”

[from researcher interviews] "I need to organize this mess.”

the term pack rat is invariably a pejorative
principles & practices: back up stuff

principle: Back up stuff to minimize unintentional loss.

[re: 13,000 email messages that participant has saved intentionally] “And they’re all stored in here. On the computer... Never have [backed them up]”

[from researcher interviews] “Unfortunately I use a lot of data that is very very big, gigabytes of stuff... and it's not backed up. It's a bad situation. But what can you do?”
principles & practices: replaceability

principle: Anything you get from the Web can be easily replaced.

“I mean nothing on here is really all that important to me, because it’s all things that I could download again if I lost it.”

“if I Google stuff, I could find these things again.”

“My pictures and my documents are more important. Because music you could always go and buy. Or you could always go and burn it somewhere else.”
so challenge 1 is assessing value and establishing provenance
A skeptical reviewer’s comment

“Seriously: what’s the hangup? As long as I take out the photos and look at them every decade or so, it’s a piece of cake. We buy a new computer every few years, spend a few minutes moving our documents folder to the new machine, we’re done. You aren’t suggesting that, come 2054, nobody will remember how JPEG works?”
**challenge 2: distributed assets**

stuff is **distributed** on and offline, on various digital media, old computers, multiple household computers, online (on Internet-based servers), on other people’s computers...

e.g. offline, possibly on outdated media

“I mean, they [Jaz drives] were new for, like, awhile, but then all of the sudden, you could write on CDs, so then Jaz dropped out of the picture. It was almost overnight.”

e.g. as email attachments

“I save everything [in email]. I never delete because I figure it’s kind of an online journal, it’s a time capsule.”
Why does this happen? (a short, incomplete list of motivations)

- informal backup
- sharing stuff with others
- using files on different computers/devices
- using network resources and services
- ...

and it’s not going to stop happening if there’s a centralized archive!
sometimes files are stored offline for a reason...

"Oh, that—that's the hard drive from my first marriage."
a performance artist’s digital stuff…

she lives in a 250 sq ft studio apartment – how far can her stuff go?
her DV camera; videos of her godsons are on DV tape. Also videos she's made of the TV

she had the DV content on the old hard drive, but not its replacement

the old hard drive used to be installed on her PC. It still probably has her old files

her old hard drive is installed in a friend's computer.

a friend maintains another website that contains her manipulated pictures of Christian Bale

she moved the novel to her PC, but the formatting got lost

she also uses Bale photos in scrapbooks

she has the novel to Rick and he printed it out for her

her Mac (not working right now) is where she's input her novel

she has a DVD burner; some of her files are on DVD (with help)

she'd like to put the files on this DVD back on the disk, but some don't open.

her website is maintained by her friend Tim, but she contributes to it and downloads photos of herself from it.

the music she creates is stored on a friend's computer. She doesn't have a copy on her computer.

she has mail on several services including Yahoo, AOL, and Hotmail. Some subset of her Favorites are on AOL. Shares photos by mail.

her website is maintained by her friend Tim, but she contributes to it and downloads photos of herself from it.
So what happens with a less naïve user and social media websites in the mix?
1.) school website, 2.) blogspot, 3.) wordpres.com (free blog host, different from wordpres.org), 4.) flickr, 5.) zoomr (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] Cathy 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] *** says: oh right.. youtube because people always tell me that they don't feel like downloading my quicktime files from archive.org
5 copies of a student animation

downloaded 387 times
3,869 views, ★ ★ ★ ★ ★
45 views, no “likes”

viewed 245 times
“really nice vid here, i enjoyed this one a lot.”
people start losing track of where everything is...
copies diverge...
added metadata gets lost (or isn’t recreated)...
resolution of photos changes...
so challenge 2 is distributed storage
A skeptical reviewer’s comment

“Seriously: what’s the hangup? As long as I take out the photos and look at them every decade or so, it’s a piece of cake. We buy a new computer every few years, spend a few minutes moving our documents folder to the new machine, we’re done. You aren’t suggesting that, come 2054, nobody will remember how JPEG works?”
But it’s not really a piece of cake.

It’s hard.

And here’s why...
scale: it’s no longer a matter of “taking out and looking at” 29 photos
we start with an unholy mix of consumer attitudes

**optimism** about the incorruptibility of digital forms

“They’re all digital files, why would they stop working?”

**fatalism** about the reliability of digital technology

“I mean, if we would’ve had a fire, you just move on.”

**fear** about vulnerability of networked digital storage

“I don’t know if I’d want to [have my] artwork, letters I read at my mother’s funeral [online]… I feel more private about that than my money.”

“128 [bit] encryption, yeah. We’d have at least that much [to protect our online photos]… 64 bits has been hacked easy.”
a brief aside about consumers, fear, and security...

the best analogy is pesticides...

c.f. consumers, pesticides, and Frierson Lake, a small lake in East Texas
...add in aggregated snafus...

all consumer study participants had registry issues, partially installed software, inexplicable dialog boxes...an aggregation of minor problems

“there’s this thing that comes up – and it’s ‘skins file’. You can’t open it; you can’t delete it; so all you can do is ‘x’ out of it to get on to whatever you’re doing.”

“I don’t know why [the media player] stopped working, just to mess with me”
and (in some cases) incomplete models of how computers work

“Kodak Memory Albums. I’m not sure if our photos are here, or Adobe. [clicks to open the app. See photo.] Okay. Nothing.”

“That’s not a photo; that’s a game.”
Viruses, spyware and malware are common – consumers are unsure how they’ve become infected or what to do.

“The conundrum that I’m in is like in order to back anything up on this computer, the computer has to be working well, and in order to get the computer working well, I should have backed up everything on this computer. D’ya know what I’m saying?”
people don’t want to expend a lot of effort for downstream return
e.g. file system organization and media labels aren’t designed for long term use

“It’s kind of weird but with some of these CDs you can tell how much is written on it by looking.”

“I have a lot of backup here from my office when I retired... I get calls from them and they want to know something. ... Ooooh! Jimi Hendrix is in there... See, this is the thing—I don’t know what—so these are all of our, uh, software. And I’m sure that Turbo Tax [with our tax returns] should be in here.”
Home users rely on friends and family for IT help. Ad-hoc support isn’t always around. Worse yet, multiple IT people may come into conflict:

“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], y’know, Jack will just do whatever he wants.”
and people rely on other people for more than IT...

Information management is a communal affair

“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.”
But these examples are drawn from the consumer study...

what about more computer-savvy people?
It’s still a problem… a slightly different problem, but still a problem.

“The problem is that, this data I have all over the place. It's very hard to remember a year later exactly where did you put that file.”

Remember that website maintainers lost their stuff by not doing anything!
the case of the disappearing podcasts

“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... and in blipmedia's google help group it was only when people ASKED HIM ABOUT IT that he explained.”
so challenge 3 is stewardship (the care of digital data)
A skeptical reviewer’s comment

“Seriously: what’s the hangup? As long as I take out the photos and look at them every decade or so, it’s a piece of cake. We buy a new computer every few years, spend a few minutes moving our documents folder to the new machine, we’re done. You aren’t suggesting that, come 2054, nobody will remember how JPEG works?”
**Long term access** is a different problem than desktop search (its closest cousin).

Like desktop search, you’re looking for a known item; unlike desktop search, you may have forgotten critical features and context.

Re-encountering may be more important than search for reclaiming forgotten material.

And remember those copies we were talking about earlier?
Web search is often a matter of satisfying...

on the Internet, any results will do
"I like doing Google searches on people I meet. And I collected some information and I guess I emailed this to her."

in fact, I just want an answer - any answer - to my question
"They’ll say, ‘okay, for Groundhog Day’ – then they’ll ask an obscure Groundhog Day question. Like, what does he eat? I never knew Punxsutawney Bill—Phil—ate a specific thing, which I can’t even remember any more … I like Google. I think it’s a really good search engine. And if not, I just Ask Jeeves. Life is too short. Because I don’t want to have 5 million choices to go through.”
Re-encountering is where the item itself reminds you of where and when you got it and why you kept it.

When I’m old and gray, this copy of *High Life* will remind me of my backpacking trip to Amsterdam “where everything’s allowed.” I’ll put it in the steamer trunk in the guest room closet...
But re-encountering must be approached with care…

“Oh, it’s looking at all the hard disk. ... [Clicks on a photo.] Ooop! Sorry! I’m ready to commit suicide.”

“I had a lot of other pictures of me similar to the one that you saw ... not pornographic but a little bit kinda, you know. Pictures like that.”

“I have, umm, erotic photos which every man downloads.”

“Now I have my 18 year old son here... And I told him, ‘Jack, you better—probably there are some porn sites on there—and do you want these ladies to see them?’”
Can you search for something you don’t remember you have?

It’s easy to forget individual items; It’s easy to forget external storage; It’s easy to forget mobile devices; — and — it’s possible to forget all of them!
<table>
<thead>
<tr>
<th>Program</th>
<th>Years</th>
<th>How I kept the email</th>
<th>Accessible?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Laurel</td>
<td>1981-1983</td>
<td>On Alto removable disk</td>
<td>No. Can’t even read the storage media.</td>
</tr>
<tr>
<td>Lafite</td>
<td>1983-1989</td>
<td>On paper</td>
<td>Yes. Printed &amp; stored in two large 3-ring binders; reread many times.</td>
</tr>
<tr>
<td>Elm</td>
<td>1994-now</td>
<td>On a file server at Texas A&amp;M</td>
<td>Yes. Still have account and access to the email software.</td>
</tr>
<tr>
<td>Eudora</td>
<td>1996-1999</td>
<td>On the original computer’s local disk</td>
<td>The hard drive on this Mac doesn’t spin up anymore. (But later found files)</td>
</tr>
<tr>
<td>Outlook (Xerox)</td>
<td>1997-1998</td>
<td>On the original computer’s local disk</td>
<td>No. I no longer have access to this computer, but it may still be in use.</td>
</tr>
<tr>
<td>Outlook (FXPAL)</td>
<td>1998-2000</td>
<td>On an in-use computer in my home</td>
<td>Yes. I used a utility to remove the password from the .pst file.</td>
</tr>
<tr>
<td>goAmerica email</td>
<td>2000</td>
<td>On the device, backed up to PC at work</td>
<td>Yes. From recovered from backup files.</td>
</tr>
<tr>
<td>Yahoo mail</td>
<td>1999-now</td>
<td>On Yahoo’s server</td>
<td>Yes. But no easy way to save them locally.</td>
</tr>
<tr>
<td>Outlook (MS)</td>
<td>2000 on</td>
<td>Server and locally on laptop</td>
<td>Yes. But it’s against company policy</td>
</tr>
</tbody>
</table>
filing sometimes = forgetting
The trouble with copies

Which Twin has the Toni?

(and which has the TV beauty shop war? See answer below.)
t1: big photo shoot

- Photos moved to desktop; some edited in Photoshop
- Photos emailed to Tim to upload to her website
- Photos written to DVD before new drive is installed

- Photos restored to new hard drive (from DVD & from website)
- Photos re-edited
- Photos attached to email to use for online dating

- How many copies does she have?
how many copies? where are they?
which have been edited? which are high res?

Original on camera flash 126-2162_IMG.jpg
File on old desktop hard drive 126-2162_IMG.jpg
File edited in photoshop Eden20.psd
File in “sent” mail (sent to art partner) Eden20.psd
File uploaded to web site (mediated) Eden20.jpg
File written to CD (mediated) Eden20.psd & 126-2162.jpg
Files restored from CD to new drive Eden20.psd & 126-2162.jpg
File downloaded from website because psd files won’t open EB.jpg
Files edited in photo-editing app EB-4U.jpg
File in “sent” mail EB-4U.jpg

Answer: at least 12 copies; 2 formats; 4 filenames; 6 file systems; and 3 resolutions (camera, web, email)
so challenge 4 is long-term access*

*of forgotten stuff of near-duplicates of misremembered stuff
whaddya trying to do here, boil the ocean?
addressing the four challenges: choosing tractable problems

- Develop techniques to assess item value and maintain item provenance
- Support distributed storage
- Provide curatorial tools and services
- Investigate new methods for long-term re-encounter and access
additional social and technical questions

- long-term value of new digital genres
  - e.g. blogs, podcasts, YouTube snippets, myspace pages, facebook profiles, and more—the stuff people have today.

- secure online services and stores
  - e.g. online banking, other financial services, medical records

- DRM-related issues

- trust and security trade-offs
  - e.g. keeping track of encryption keys and passwords

- ‘traditional’ digital preservation questions
  - e.g. developing format registries; emulation services
the other thing to remember is that it’ll take a village...

this problem calls for partnerships and cooperation among libraries, publishers, non-profits, software companies, social media sites, records repositories, and Internet services providers...

- develop a sense of cultural stewardship
- develop workable copyright policies
- address constraints introduced by patents and proprietary formats
- create a financially sustainable enterprise
credits

- personal digital archiving field study collaborators: Sara Bly and Francoise Brun-Cottan

- Web site recovery study collaborators: Michael Nelson and Frank McCown (ODU)

- Catharine van Ingen, the Community Information Management project at MSR SVC (Doug Terry, Ted Wobber, Tom Roddehoffer, and Rama)
questions?

contact info:

cathymar@microsoft.com
marshall@csdl.tamu.edu

http://www.csd1.tamu.edu/~marshall