Benign neglect in a digital world: a pragmatic look at personal digital archiving

Cathy Marshall
Microsoft Research, Silicon Valley

First Digital Lives Research Conference
9 February 2009
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
backwards drive. Is there a floppy drive? Do I need to buy an external floppy disk drive?

A: No CD is easy to make and store. If you're transferring files from a computer to a CD or
vice versa, you can use a CD-RW drive. CD-RW drives are becoming increasingly
popular and are available at most computer stores.

Q: Is there a way to use a CD-RW drive to create a DVD?

A: CD-RW drives are not designed to create DVDs. DVDs require a special drive designed
to write data to DVDs. However, some CD-RW drives can create DVDs, but they
will require additional software.

Q: Is there a way to use a CD-RW drive to create a Blu-ray disk?

A: CD-RW drives are not designed to create Blu-ray discs. Blu-ray discs require a
special drive designed to write data to Blu-ray discs. However, some CD-RW drives
may be able to create Blu-ray discs, but they will require additional software.

Q: Can I use a CD-RW drive to create a CD-R or CD-RW disk?

A: Yes, CD-RW drives can create both CD-R and CD-RW disks. CD-R disks can be
copied and used as many times as you like, while CD-RW disks can be rewritten.

Q: Is there a way to use a CD-RW drive to create a music CD?

A: CD-RW drives can be used to create music CDs, but you will need a special
program to create the music files and burn them to CD. Many music creation
software programs include the ability to burn CDs.

Q: Is there a way to use a CD-RW drive to create a data CD?

A: Yes, CD-RW drives can be used to create data CDs. Data CDs can be
written to and read from, and are useful for storing and transferring
large amounts of data.

Q: Is there a way to use a CD-RW drive to create a video CD?

A: CD-RW drives are not designed to create video CDs. Video CDs require a
special drive designed to write data to video CDs. However, some CD-RW drives
may be able to create video CDs, but they will require additional software.

Q: Is there a way to use a CD-RW drive to create a photo CD?

A: CD-RW drives can be used to create photo CDs, but you will need a special
program to create the photo files and burn them to CD. Many photo creation
software programs include the ability to burn CDs.

Q: Is there a way to use a CD-RW drive to create a document CD?

A: Yes, CD-RW drives can be used to create document CDs. Document CDs can
be written to and read from, and are useful for storing and transferring
large amounts of files.

Q: Is there a way to use a CD-RW drive to create a movie CD?

A: CD-RW drives are not designed to create movie CDs. Movie CDs require a
special drive designed to write data to movie CDs. However, some CD-RW drives
may be able to create movie CDs, but they will require additional software.

Q: Is there a way to use a CD-RW drive to create a music CD, a data CD, a video CD, a
photo CD, a document CD, and a movie CD?

A: Yes, CD-RW drives can be used to create all of the above types of CDs.
However, you will need to use different programs to create each type of CD.

Q: Is there a way to use a CD-RW drive to create a DVD, a Blu-ray disc, a music CD, a
data CD, a video CD, a photo CD, a document CD, and a movie CD?

A: Yes, CD-RW drives can be used to create all of the above types of discs.
However, you will need to use different programs to create each type of disc.

Q: Is there a way to use a CD-RW drive to create a CD-R or CD-RW disk, a DVD, a
Blu-ray disc, a music CD, a data CD, a video CD, a photo CD, a document CD,
and a movie CD?

A: Yes, CD-RW drives can be used to create all of the above types of discs.
However, you will need to use different programs to create each type of disc.

Q: Is there a way to use a CD-RW drive to create a CD-R or CD-RW disk, a DVD, a
Blu-ray disc, a music CD, a data CD, a video CD, a photo CD, a document CD,
and a movie CD, but with a CD-RW drive that does not support rewritable modes?

A: Yes, CD-RW drives can be used to create all of the above types of discs.
However, you will need to use different programs to create each type of disc.
“…neglect can sometimes be an artifact’s best friend.”

- G. Thomas Tanselle
“Statement on the Significance of Primary Records”
What exactly is a home page? In the simplest terms, it 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.

Home on the Net, John Seabrook, 16 October 1995
Apple QuickTake digital camera (circa ‘95)
My 1995 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 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?”

Jeff Rothenberg, “Ensuring the Longevity of Digital Documents”
SCIAM, Jan ‘95
“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.”
fast forward to 2009
there are more than 3.2 billion personal photos on Flickr

and if that’s not enough, Facebook has AT LEAST twice that many...
Two familiar solutions: (1) shove everything into a big database in the cloud and decode it later (the infinite U-Store-It approach) ...

...or (2) safe storage and self-describing digital objects
But wait!

Are we still solving the same problem that Rothenberg was addressing in 1995?
How can we uncover assumptions and pose new technical solutions?

by paying attention to the whole social/technical sphere...
real people, real practices, real technology

bringing together the results of different studies
4 challenges: a capsule summary

- people are accumulating stuff—some that matters to them, some that doesn’t—and it’s only going to get worse.
- people try to keep their stuff safe via *ad hoc* distribution and replication.
- people approach digital stewardship with the same level of benign neglect, good intentions, and contradiction that they bring to other aspects of their lives.
- retrieval from a long term store is unlikely to be like googling or desktop search.
challenge #1: accumulation, value, and provenance
When asked when he ever got rid of digital stuff, one person I interviewed 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.”
challenge #2: ad hoc replication as safety net
[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] Cathy Marshall 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

so people put copies of their stuff in different places for different reasons.

and safety is an important side effect!
We think of the local copy as archival (and it is in the sense that it’s highest fidelity)

“The good thing about the photos is that there’s always an intermediary step. I mean, like 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. So that’s why I don’t care so much about Flickr evaporating.”

But... the web copies have been augmented with useful organization and metadata (e.g. tags and possibly comments and ratings)

“I didn’t lose the pictures, but I was sorry that I had lost the collections and the organization, and you know. I’m sure I have the pictures somewhere still. But fishing them out and recreating it was not feasible.”
we attribute loss to technological catastrophes, but often it isn’t
For scholars, the key vulnerability is changing organizations; it is more cataclysmic than technology failures.

Sources of unintentional loss

- files are misplaced in the shuffle
- accounts evaporate more suddenly than expected
- infrastructure changes
- replication schemes are re-centralized

“When you change jobs, you typically lose a lot of things. So my life starts in 2001.”
challenge #3:

the people we’re talking about don’t want to spend very much time (or very much money) on digital curation
“It’s funny though.... 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.”
“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.”

“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?”

“It’s kind of weird but with some of these CDs you can tell how much is written on it by looking.”
“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.”
It’s easier to *keep* than to *cull*, but it’s easier to *lose* than *maintain*. 
challenge #4:

retrieval from long-term storage isn’t like web search;
it isn’t like desktop search;
it isn’t like browsing the file system...
people may not remember what they have and they make both kinds of errors.

Even if they remember what they have, the context is vague. Often they’ve forgotten where they put it (especially given the degree of distributed storage that is typically used)

Finally, there’s the problem of coming up with the the ‘good’ copy/the original/ground truth.
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.

Copy of *High Life* reminds informant of her backpacking trip to Amsterdam “where everything’s allowed.” She stows it in the steamer trunk in the guest room closet with other high-value emotionally evocative items.
t1: big photo shoot

- t2: photo moved to desktop & edited in Photoshop
- t3: photo emailed to Tim to upload to her website
- t4: photo written to DVD so new drive can be installed

- t5: Photo restored to new hard drive (from DVD, then from web site)
- t6: photo re-edited in other app
- t7: photo 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)
let's recap:

accumulation
distribution
curation
access
Notice these four challenges translate to four key questions you'd need to answer to build a personal archive:

1. what should we keep?

2. where should we put it?

3. how should we maintain it? and

4. how will we ever find it again?
each question suggests part of an architecture... but that’s part 2, and we’re out of time today
Most important lesson: No single solution! No central store!

- catalog-driven

- may involve other storage sites
  - e.g. free software to create S3 backup
  - e.g. for-pay “vault” software

- may involve different methods for handling medium- and low-value items

- any solution should acknowledge the human tendency toward benign neglect
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, Rama, and Rama Kotla)
questions

contact info:
cathymar@microsoft.com
http://www.csdl.tamu.edu/~marshall
http://research.microsoft.com/~cathymar