The Sustainability of Everyday Stuff: A pragmatic look at personal digital archiving

Cathy Marshall
Microsoft Research, Silicon Valley

Inaugural Symposium, UT iSchool
30 October 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 CD-RW drive that you can plug into a floppy drive? Do I need a special program to read files on a CD-RW drive? Do I need to change the CD-RW drive's letter assignment? A: The CD-RW drive is not compatible with the floppy drive. You will need a special program to read files on a CD-RW drive. You will also need to change the drive's letter assignment.

Q: I recently switched from America Online to Comcast. With AOL I could use Microsoft Outlook. Is there a similar email program available? A: There is a similar email program available. It is called Express Mail and it is available for free.

Q: I want to have a PowerPoint presentation that shows side-by-side views of two slides. Is there a way to do that? A: Yes, you can use PowerPoint's side-by-side view feature to show two slides at once.

Q: How do I set up a network printer? A: You will need to install the printer software and connect the printer to your computer via a network cable.

TIP OF THE WEEK

In a recent column, Microsoft announced a new feature that allows users to easily print to a network printer. The feature is available in Windows 7 and later versions of Windows.

Saving file with a CD-RW drive

David Einstein
Computing Q&A

would not require a steep learning curve for a relatively straightforward set of things to track. Would you ever have to do anything more complicated than this? A: It is possible to do more complicated tasks, but it would require a significant amount of effort.

Microsoft Word

A: Here's the deal. AOL has its own spell-checker. But when I switched to Outlook, I had to use Outlook's spell-checker. I'm not sure if Outlook's spell-checker is any better than AOL's. But if you don't have Outlook, you can use Microsoft Word's spell-checker. It's not as good as Outlook's, but it's still pretty good.
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.

participant in an interview study who had lost his personal and business websites in a crash
“...neglect can sometimes be an artifact’s best friend.”

- G. Thomas Tanselle

“Statement on the Significance of Primary Records”
Multiple copies of a rare vinyl record... reel-to-reel tape used to archive the rare vinyl record...
The United States of America [EXTRA TRACKS]

List Price: $17.98
Price: $17.98 & eligible for FREE Super Saver Shipping on orders over $25. Details

Want it delivered Wednesday, February 27? Order it in the next 20 hours and 22 minutes, and choose One-Day Shipping at checkout. See details

44 used & new available from $11.29
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
how quaint!

how last millenium!
Apple QuickTake digital camera (circa ‘95)
my trip to Graceland

resulted in 29 mostly awful photos in tiff format...
a call to arms circa 1995

“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.”

Figure 9: Using emulation to read an obsolete digital document
fast forward to 2009
there are > 4 billion personal photos on Flickr

but that’s just the tip of the iceberg...
Facebook has AT LEAST three times that many (15B/60B) and Image Shack has AT LEAST four times that many (20B)
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?
personal digital archiving

archiving a personal digital collection
Michael Joyce archiving his own digital stuff

Gabby Redwine archiving Michael Joyce’s digital stuff
How can we uncover hidden assumptions and pose new technical/social solutions?

We need to pay attention to all of it: look at the whole social/technical sphere...

- digital stuff
- social interaction
- technology
- security and access control
- storage & retrieval: platforms and repositories
- format/self-describing digital objects

...and other aspects...
This talk draws on multiple data sources

- consumer field study in 3 cities
  
  what people save, where they keep it, and are their strategies working?

- survey and interviews of people who were trying to recover 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
Do people believe they archive their digital stuff? (benign neglect realized)

- 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
personal digital archiving:

4 challenges & themes
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 other IR problems
challenge #1: we simply accumulate content without regard for its value or provenance
When asked when he ever got rid of digital stuff, one person answered,

“[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.”

When asked when he ever got rid of digital stuff, one person answered,

“Yes, but not in any systematic manner. ... It’s more like, I have things littering the desktop and at some point it becomes un navigable...

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.”
Loss as a means of culling collections

“If [my email] 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.”[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?”
It’s easier to *keep* than to *cull*.
So what’s the problem? Storage is cheap and getting cheaper. Why not just keep EVERYTHING?
why might you want to keep everything?

- It’s difficult to predict an item’s future worth.
- Deletion is hard, thankless work.
- Filtering and searching can readily locate the gems among the gravel.
If you’re using your archive as a true memory prosthesis, deleting defeats the whole purpose...
Let's take a look at personal scholarly archives as a concrete example.

**Now:** “I'll probably keep [the reviews] forever. As well as my replies and things like that.”

**20 years later:** the archive contains
- PDFs of publications
- Some bibliographic resources
it doesn’t contain (but should possibly)
- Datasets (in some domains they clearly do)

**what happened?**
in the end, there are tradeoffs...

- storage is cheap, but storage is only part of the equation. (e.g. human attention)

- there are *legal* and *emotional* constraints.

- Most of us (sadly) don’t have Vicki
challenge #2: we use 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.) 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 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.

data safety is an important side effect!
This leads people to use circular reasoning about data safety and where ‘truth’ lies...

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)

“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.”
the distributed copies are at different fidelities and have different metadata in fact... each copy takes on a life of its own.

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

viewed 245 times
“really nice vid here, i enjoyed this one a lot.”
we attribute loss to purely technological catastrophes, but it often isn’t
[Closed] My account deleted!

Shéhérazade (vanished... killed by Flickr staff) says: name reply, icon reply

Flickr staff deleted my account without any reason and warning.
All my pics was taken by me and flagged as restricted
in respect of the community guidelines.
150,000 visits in six months,
22 testimonials,
200 comments for each image.
One of the most famous and respected streams
in the whole Flickr.
ALL VANISHED without reason!!!
I mailed to ask why they destroyed all my work,
and TERRENCE replied saying that
i posted photos not taken by me.
IT'S NOT TRUE AT ALL!!!
I took ALL MY PICS.

I want my account back! I paid for it!

Posted at 6:13PM, 25 January 2009 PST (permalink)
heather (staff) edited this topic 4 hours ago.
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: people don’t want expend much effort on digital curation (nor do they have wherewithal or skills)
“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?**”
scale: it’s no longer a matter of “taking out and looking at” 29 photos
people may well be
better at capturing
better at creating
better at shaping
better at sharing
“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 [that] 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.”

“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 a textbook IR problem
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.”
memory is vulnerable to errors of both sorts

distributed storage (sometimes offline, sometimes in walled gardens) makes retrieval more difficult

the reference copy may not be in a predictable place
physical personal archives rely on re-encounter

*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.
Which Twin has the Toni?

(and which has the T15 beauty shop hair? See answer below.)
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 DVD (mediated) Eden20.psd & 126-2162.jpg
Files restored from DVD 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)
When she searches, which copy will she find first?

will she know that there are other ones?

will she understand the differences among them?
let's recap:

→ accumulation
→ distribution
→ curation
→ access

Esther Dyson in her office (photo courtesy of Ramana Rao)
whaddya trying to do here, boil the ocean?
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. who should maintain it and how? and
4. how will we ever find it again?
Part 2: each question suggests part of an architecture
what should we keep?
value dictates how many copies you'd want and how you'd store them

- **known high-value stuff**
- **medium value stuff**—we used to rely on benign neglect to leave us w/some of it
  
  Preservation through use—*the more I use these items, the more valuable they’re likely to be and the better their chance of survival*

- **lower value stuff**—ambivalent attitude

- **the controversial stuff**—deleted? are you sure? (i.e., it may be more important to make sure something’s gone...)
## Use-based heuristics for assessing value

<table>
<thead>
<tr>
<th>type</th>
<th>value indicator</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>source</td>
<td>created locally</td>
<td>novel (.doc file)</td>
</tr>
<tr>
<td></td>
<td>p-t-p file</td>
<td>bootlegged music (.mp3 file)</td>
</tr>
<tr>
<td>action</td>
<td>edit metadata</td>
<td>name a photo</td>
</tr>
<tr>
<td></td>
<td>view content</td>
<td>play a song</td>
</tr>
<tr>
<td>disposition</td>
<td>upload to service</td>
<td>share on Flickr</td>
</tr>
<tr>
<td></td>
<td>remove</td>
<td>drag to trash</td>
</tr>
</tbody>
</table>
where should we put it?
personal archiving can be an application on top of a synchronization infrastructure.

- **union content catalog**
  - where the item is stored
  - assessed or asserted value
  - composite metadata (e.g. from social media sites)
  - provenance (e.g. which is the reference copy)

- **device/repository manager/kb**
  - characteristics (e.g. reliability, capacity)
  - access management
  - ping service

- **retrieval/re-encounter machinery**
an architecture for storing high-value items
who should maintain it and how?
invisible routinized activities that can be automated as services
- virus and malware checks (68% of consumers had virus/malware infections in 2005)
- identify at-risk files (photos in RAW format)
- refresh storage media

communal organizing and labeling activities (harnessing the power of social networks)
- within social network: tags, narrative, etc.
- records of activity (to gauge popularity, e.g.)
- on an institutional level, format registries

stuff requiring human intervention
- an important role for iSchools and libraries
- training consumers
- providing institutional infrastructure
how will we ever find it again?
we’ve forgotten it altogether: methods for re-encounter

we’ve got some context and a rough idea of what we want: faceted browsing

we’ve got a good idea of what we want: alternative visualizations (and desktop search)
Re-encounter is probably more effective if the item is either in-context (i.e. IQ-based) or high-value (browser-based).
techniques for re-encounter

- stable personal geography
  - differentiated places

- value-based organization
  - re-encounter of high-value items

- better presentation of item surrogates
  - develop good reduced representations of media types other than photos!

(implicit query)
But re-encountering techniques must be approached with care...

“Oh, it’s looking at all the hard disk. ... [Clicks on a photo.] Ooops! 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?’”
we’ve got some context and a rough idea of what we’re looking for: faceted browsing (from myLifeBits)
alternative presentations: annotated time line (also from myLifeBits)
browsing tools should help you find and select among copies and near copies; they should help you sort out matters of provenance...
What's the message?
No single solution!

- catalog driven
- probably involves other storage sites
  - e.g. free software to create S3 backup
  - e.g. for-pay “vault” software
- use different methods for handling medium- and low-value items
- new institutions/outreach/cultural expectations
- acknowledge the power of benign neglect and the need to forget!
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 R., and Rama Kotla)
any questions?

contact info:

cathymar@microsoft.com

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

blog—http://ccmarshall.blogspot.com
twitter—http://twitter.com/ccmarshall