people, their digital stuff, and time: opportunities, challenges, and life-logging Barbie

Cathy Marshall
Microsoft Research, Silicon Valley
http://research.microsoft.com/~cathymar

Code4Lib
23 February 2010
I visited the Internet Archive last week...
Brewster Kahle said that the Web still fits compactly in a 19’x 8’ x 8’ shipping container... and that, incidentally, a Web page weighs in at around 80 micrograms
let's turn back the clock to 1995.
that so much of the stuff on the Web is personal media is a relatively new story...

i.e. little did we know 15 years ago that the exoticism of wearable computing research would be realized as life-logging Barbie
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
born-digital assets circa 1995: for me, 29 photos of Graceland taken with an Apple Quicktake camera
today, there are more than \(4.3\) billion personal photos on Flickr

Photobucket has at least 2x that (>7B)  
Facebook has at least 3x that (>15B/60B)  
and Image Shack has at least 4x that (>20B)
3—or perhaps 4—things to think about when we mix people, their stuff, and time

some ruminations about personal digital archiving derived from 20—or perhaps 25—years of feral ethnography and real studies
let’s ruminate!
Thing 1: people rely on benign neglect as a *de facto* stewardship technique & collection policy
kid: “They should just save Facebook. That is our generation’s scrapbook, yearbook, Guinness World Record...”

kid2: “But obviously you can’t save everything”

LOC narrator: “But the truth is, digital information will survive only as long as someone takes care of it.”
personal digital archiving

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

Gabby Redwine archiving Michael Joyce’s digital stuff
stewardship...
for everyday people, the road to digital stewardship is paved with good intentions
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

6 months later, he still doesn’t back up his stuff!
“…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...
the same record on Amazon today, courtesy of the ‘long tail’ phenomenon.
benign neglect

*yes. I could knit a complete second cat with the stray fur from the first one

as collection policy
the mean girls at their table in the junior high cafeteria

postcard from a friend on an archeological dig in Greece
“[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.”
Is that really a problem? Storage is cheap and getting cheaper. Why not just keep EVERYTHING?
let’s take a closer look at the *keep everything collection policy*
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 your archive acts as a memory prosthesis, deletion defeats the whole purpose...
It’s easier to *keep* than to *cull*
“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?”
Now: “I'll probably keep [the reviews for my papers] forever. As well as my replies and things like that.”

20 years later: the archive contains

- PDFs of publications
- Some bibliographic resources

and that’s good enough...

- What about the datasets? Maybe someday we’ll keep them, but we don’t do it yet
It’s easier to *keep* than to *cull*,

but it’s easier to *lose* than *maintain*. 
the implication? not all long-term personal stores need to perform with the same level of reliability

known high-value stuff

medium value stuff—want 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? (ensure it’s forensically unrecoverable!)
use-based heuristics help assess 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>received via bit torrent</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>
Thing 2: no single preservation technology/
repository/
file system/
cloud store
will win the battle for your stuff...
Today, there are two standard technical solutions: (1) shove everything into a great big database in the cloud and decode it later (the Oscar Madison approach) ... 

...or (2) safe storage and self-describing digital objects (the Felix Unger approach)
[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

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

data safety is a side effect!
we attribute loss to purely technological catastrophes, but it often isn’t
the social metadata is valuable to some users!
replication and loss
in personal scholarly archives

• 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.”
instead of centralizing, we'll be knitting together stores and services

in collaboration with Catharine van Ingen
No single archive!

- catalogs knit distributed stores together
- different levels of security
- different access patterns
- new institutions/
  new cultural expectations
Thing 3: Forget about digital originals or reference copies
Which Twin has the Toni?

(and which has the P&G beauty shop wave? See answer below.)
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, captions, and comments)

“I didn’t lose the pictures, but I was sorry that I had lost the collections and the organization. I’m sure I have the pictures somewhere still. But fishing them out and recreating it was not feasible.”
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

how many copies does she have?

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
<table>
<thead>
<tr>
<th>File Type</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original on camera flash</td>
<td>126-2162_IMG.jpg</td>
</tr>
<tr>
<td>File on old desktop hard drive</td>
<td>126-2162_IMG.jpg</td>
</tr>
<tr>
<td>File edited in photoshop</td>
<td>Eden20.psd</td>
</tr>
<tr>
<td>File in “sent” mail (sent to art partner)</td>
<td>Eden20.psd</td>
</tr>
<tr>
<td>File uploaded to web site (mediated)</td>
<td>Eden20.jpg</td>
</tr>
<tr>
<td>File written to DVD (mediated)</td>
<td>Eden20.psd &amp; 126-2162.jpg</td>
</tr>
<tr>
<td>Files restored from DVD to new drive</td>
<td>Eden20.psd &amp; 126-2162.jpg</td>
</tr>
<tr>
<td>File downloaded from website because psd files won’t open</td>
<td>EB.jpg</td>
</tr>
<tr>
<td>Files edited in photo-editing app</td>
<td>EB-4U.jpg</td>
</tr>
<tr>
<td>File in “sent” mail</td>
<td>EB-4U.jpg</td>
</tr>
</tbody>
</table>

*at least 12 copies; 2 formats; 4 filenames; 6 file systems; and 3 resolutions (camera, web, email)*
16 ‘identical’ photos of a really big catfish...
each has taken on a life of its own...
<table>
<thead>
<tr>
<th>Fish Type</th>
<th>Weight</th>
<th>Location/Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vietnamese catfish</td>
<td>187.4 pounds</td>
<td>Wels catfish, a breed which can get larger</td>
</tr>
<tr>
<td>140 lb. catfish caught in</td>
<td>Photo of a</td>
<td>Giant Catfish</td>
</tr>
<tr>
<td>Lake Texoma.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A BIG Wels Catfish (187lbs)</td>
<td></td>
<td>caught by Lucas Van Der Geest</td>
</tr>
<tr>
<td>424 pound Mekong catfish</td>
<td>“Two men catch catfish by sticking their hand elbow deep into the mouth of the fish.”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“that certainly is a big fish!”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“Fishermen with Giant Catfish: real image”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Catfish from Mississippi, just shy of 646 lbs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“It's called a Wel's Catfish. They get bigger than this!!!”</td>
<td></td>
</tr>
</tbody>
</table>

Each has grown its own social metadata...
A real example: an animated music video
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.”
object (bits or surrogate)

- derived copy (content is modified)
  - near copy
  - transcoded algorithmic
    - reversible
    - lossy
  - edited manual
    - added version (keep $v_1 \ldots v_n$)
    - replaced version (keep only $v_n$)
- exact copy
  - added version (keep v$_1$…v$_n$)
  - replaced version (keep only v$_n$)
  - expert
  - authorial
  - other
- enhanced copy (metadata is added)
  - descriptive metadata
    - expert
    - authorial
    - other
  - social metadata
    - annotative
    - cumulative/aggregate
- metadata

where are the tools that’ll let me harvest the metadata the copies have grown?

where’s the search tool for gathering the copies (rather than just de-duping them)?
Speaking of search...

Thing 4: Given things 1-3, there will be some interesting opportunities to take a fresh look at searching and browsing
we're pretty blasé about search...
why searching distributed personal archives is different...

we might have forgotten it altogether: re-encounter

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

we know exactly what we want: visualizations and desktop search

and there’s the whoops factor...
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)
whaddya trying to do here, boil the ocean?
doesn’t this look like opportunity to you?
many, many **bottom up efforts**—collections, policies, tools, practices... personal archiving as a cottage industry

from the SALT project at Stanford
new institutions

SF Mayor Gavin Newsom on the Internet Archive's hiring 100 people to scan books and microfilm from the unemployment rolls leveraging a matching system using stimulus dollars.

Start at 2min 15 seconds.

http://www.youtube.com/watch?v=oaR6AURJ2UM

We are gearing up under a similar program in LA. We hope other cities will follow.

-brewster
new opportunistic uses of massed data
for analysis... (the world is my dataset)

<table>
<thead>
<tr>
<th>Tag word</th>
<th>Frequency (items w/tags)</th>
<th>Word category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milan</td>
<td>85%</td>
<td>place</td>
</tr>
<tr>
<td>Italy</td>
<td>66%</td>
<td>place</td>
</tr>
<tr>
<td>Galleria (&amp;variants)</td>
<td>26%</td>
<td>place</td>
</tr>
<tr>
<td>bull</td>
<td>25%</td>
<td>artifact</td>
</tr>
<tr>
<td>Emanuele</td>
<td>14%</td>
<td>place</td>
</tr>
<tr>
<td>Vittorio</td>
<td>13%</td>
<td>place</td>
</tr>
<tr>
<td>Europe</td>
<td>12%</td>
<td>place</td>
</tr>
<tr>
<td>200x</td>
<td>11%</td>
<td>context</td>
</tr>
<tr>
<td>travel</td>
<td>9%</td>
<td>context</td>
</tr>
<tr>
<td>luck</td>
<td>9%</td>
<td>story</td>
</tr>
</tbody>
</table>
for aggregate display (watch the hole develop as people spin on their heels)
last words...

the power of benign neglect

no single solution

the secret lives of copies

retrieval from cold storage

new opportunities lie in the aggregation of individual archives and efforts
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, Rama R., and Rama Kotla)
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