new approaches to personal archiving

Cathy Marshall
Microsoft Research, Silicon Valley

FDIS 2008, Woods Hole MA
1 July 2008

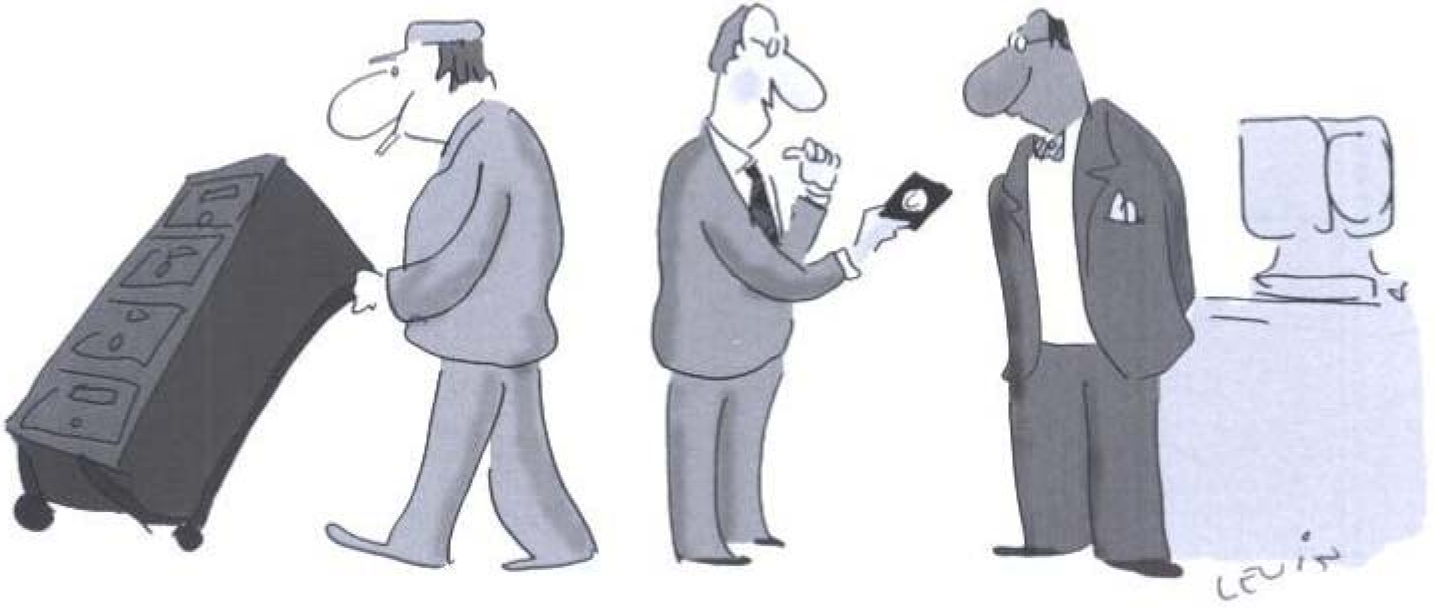
1995

from the *New Yorker* just 13 years ago!

In the simplest terms, [a home page] 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.

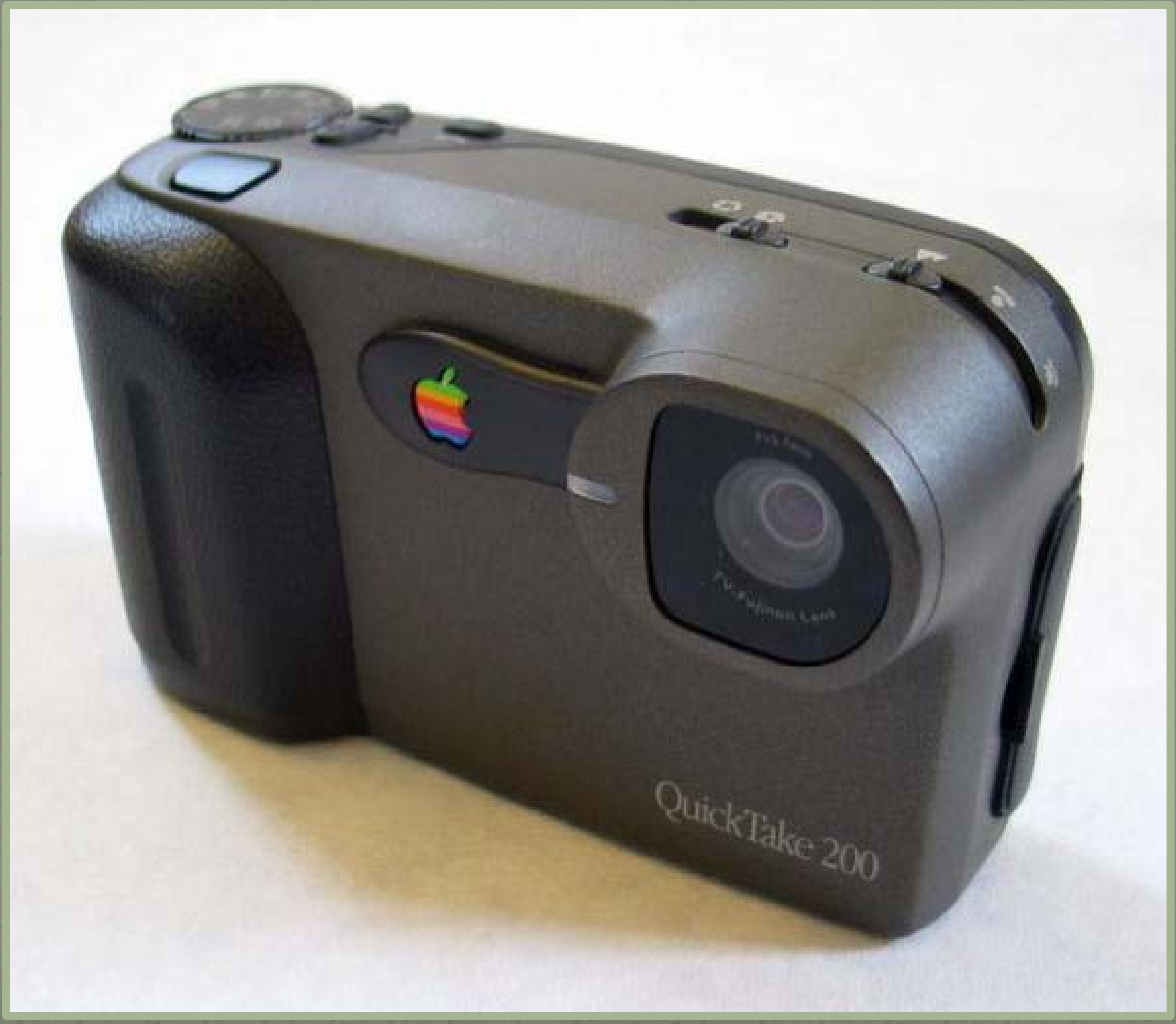
Home on the Net, John Seabrook, 16 October 1995 how quaint!

how last millenium!



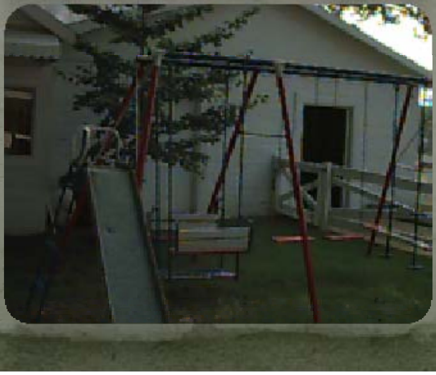
"Everything that was in that filing cabinet is now on this little disk. Except of course for my bottle of scotch."

Apple QuickTake digital camera (circa '95)



my trip to Graceland in 1995

29 mostly awful low resolution 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

…his approach: emulation

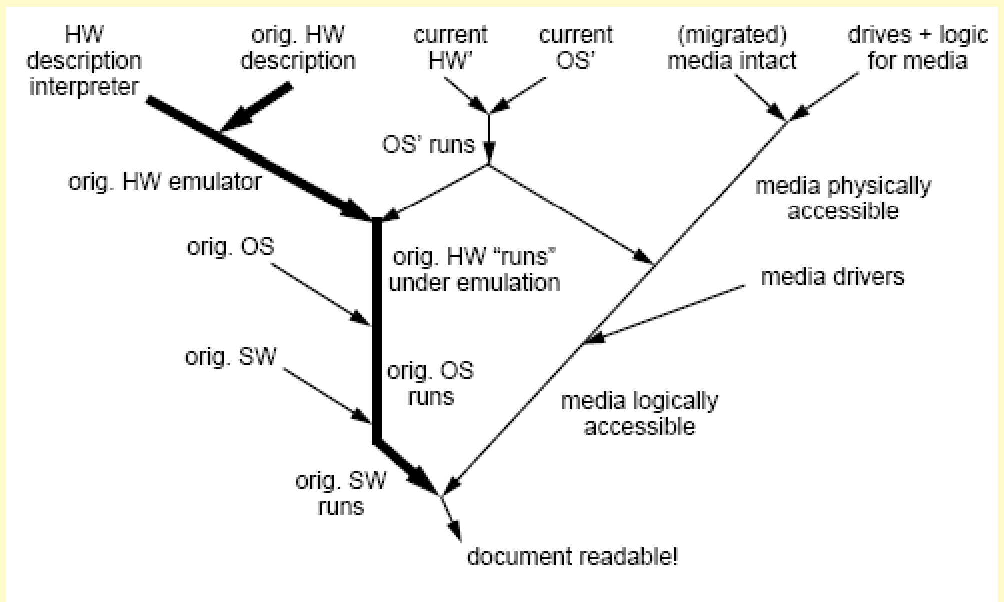


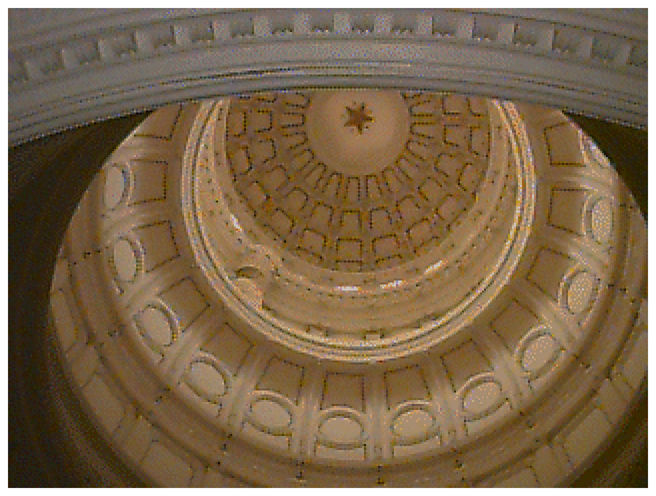
Figure 9: Using emulation to read an obsolete digital document

"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 2008

there are more than 2.6 billion personal photos on Flickr

Flickr photo # 2,626,042,804



Quicktake #26

and if that’s not enough,
Facebook has at least twice that many…

Two solutions that I've heard most often: (1) shove everything into a big database in the cloud and decode it later (aka benign neglect) …

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

…and (2) safe storage and self-describing digital objects

real people, real practices, real technology

now and down the line…

4 challenges: a capsule summary

- people are accumulating stuff that matters to them and it's only going to get worse
- people try to keep stuff safe via *ad hoc* distribution and replication
- people don't want to spend time taking care of what they've amassed (even the good stuff)
- retrieval from a long term store is unlikely to be like googling or desktop search

challenge #1:
accumulation, value, and provenance

"[when I buy a new computer] I transfer everything. … I have someone else do it for me… [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 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

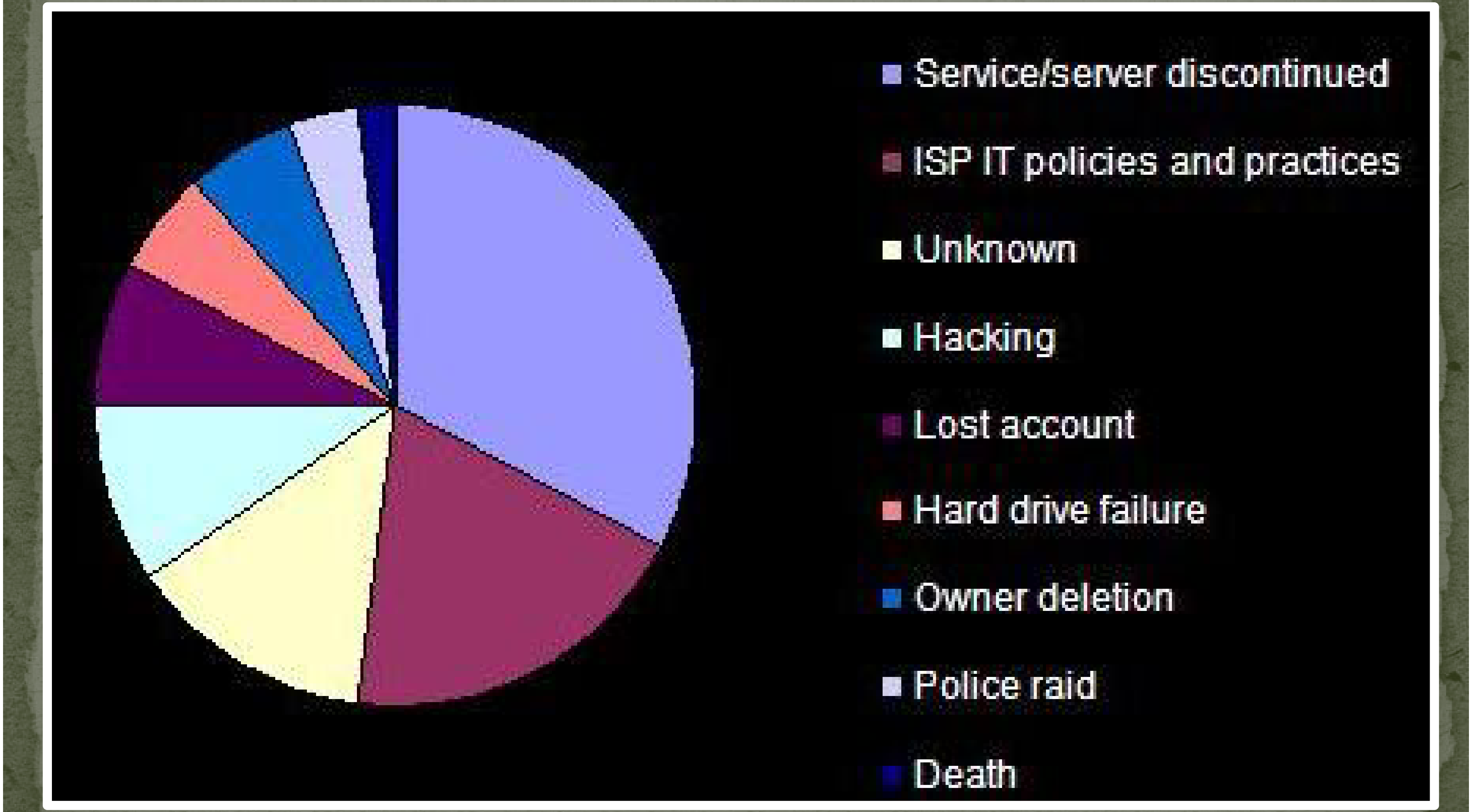
[11:09:24 PM] g says: [There are] 6 [online places where I store things] in all. 1.) school website, 2.) blogspot, 3.) wordpress.com (free blog host, different from wordpress.org), 4.) flickr, 5.) zooomr (for pictures, they offer free "pro" accounts for bloggers, but even for non-pros, they don't limit 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 they think that'll make it safe!

we attribute loss to technological catastrophes, but it often isn't



that's the thing about replication… think about it in our own personal 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
  - digital belongings become a jumble
  - *replication schemes are re-centralized*

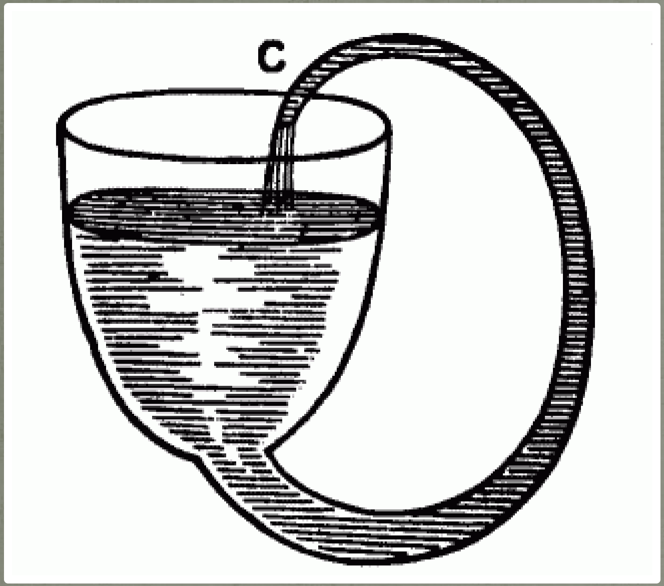
"When you change jobs, you typically lose a lot of things. So my life starts in 2001."

challenge #3:
no-one wants to spend time (or money) on digital curation

“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.”



It's easier to *keep* than to *cull*,
but it's easier to *lose* than *maintain*.

challenge #4:
retrieval from long-term storage

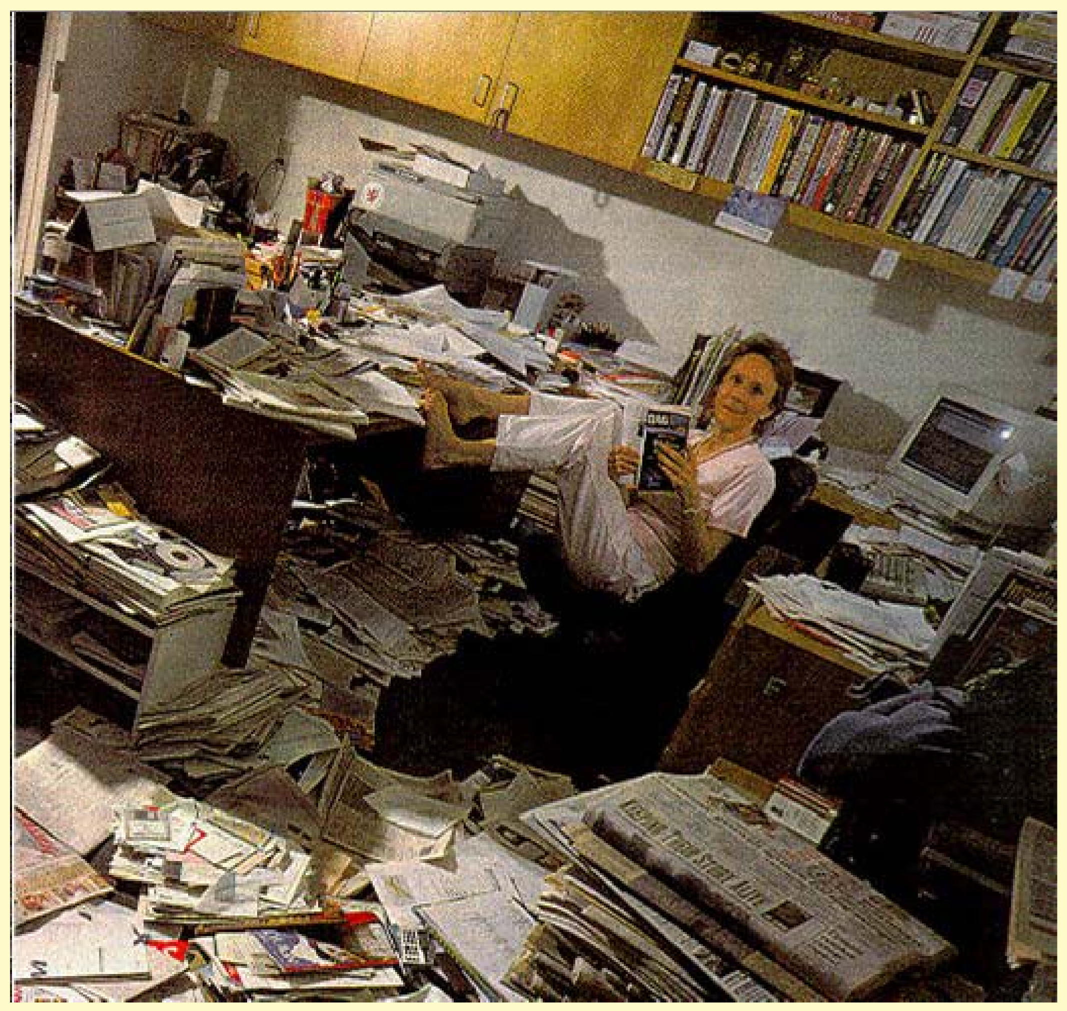
it isn’t like web search;
it isn’t like desktop search;
it isn’t like browsing the file system…

people don’t 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.

Even if they remember what they have and where they put it, they don’t remember which is the ‘good’ copy.

let's recap:



→ accumulation
→ distribution
→ curation
→ access

Esther Dyson in her office (photo courtesy of Ramana Rao's blog)

we want to build a *cheap* personal archive.

in fact… *we want it to be free*

to do this, we have to answer four questions (not necessarily in this order)

1. what should we keep?

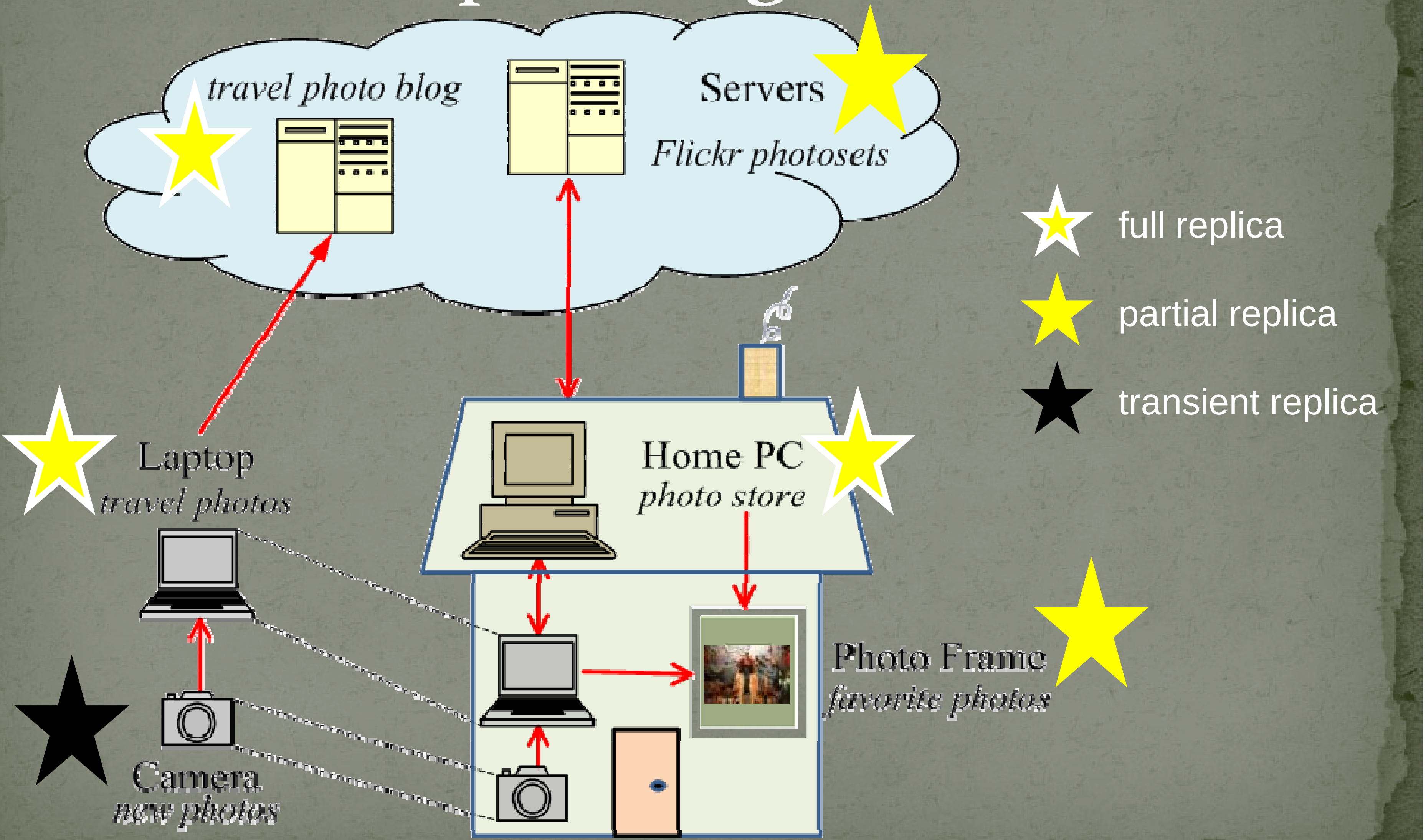
2. where should we put it?

3. how should we maintain it? and

4. how will we ever find it again?

where should we put it?

a consumer photo vignette



travel photo blog

Servers

Flickr photosets

Laptop

travel photos

Home PC

photo store

Photo Frame

favorite photos

Camera

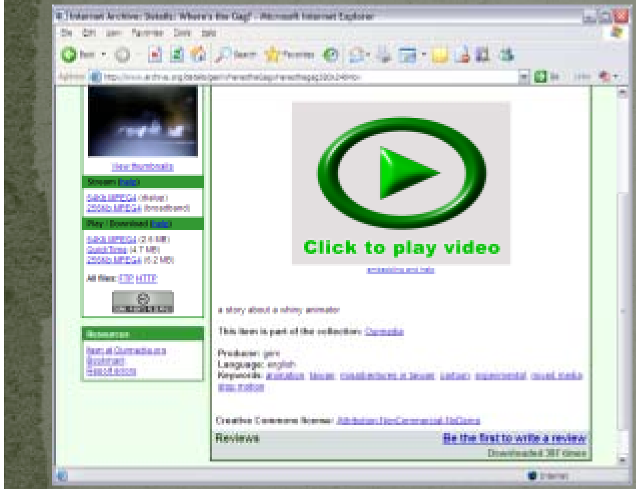
new photos

full replica

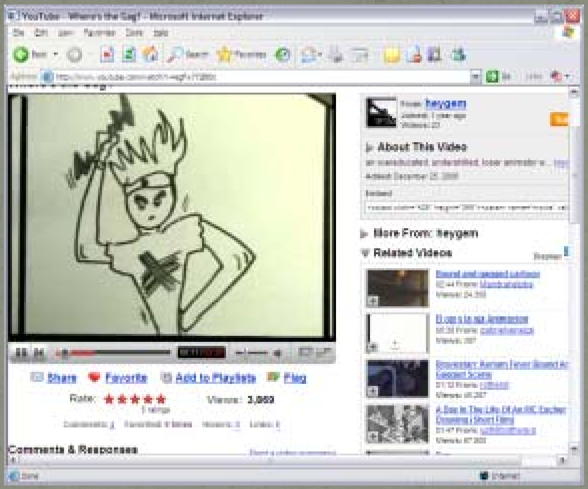
partial replica

transient replica

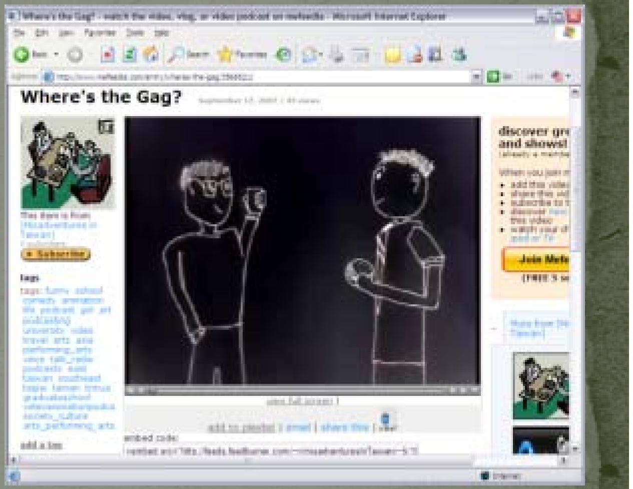
and let's not forget that it's not just content: some copies will be growing divergent metadata



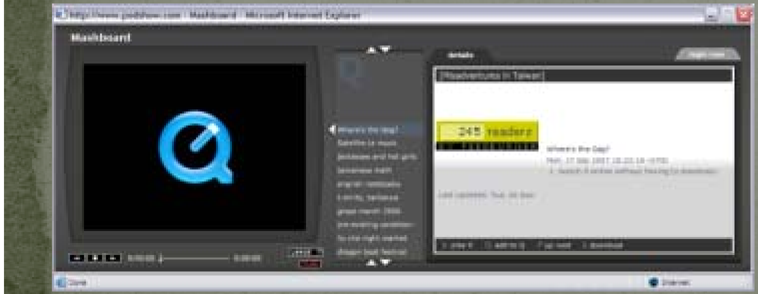
downloaded 387 times



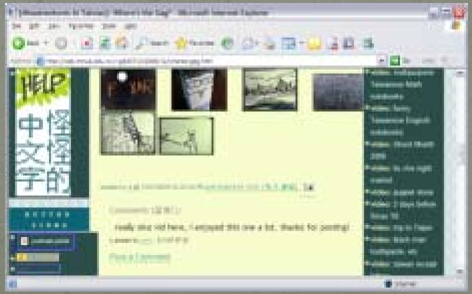
3,869 views, ★ ★ ★ ★ ★



45 views, no "likes"



viewed 245 times



"really nice vid here, i enjoyed this one a lot."

people employ circular reasoning about data safety

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 copy has been augmented with useful organization and metadata (e.g. tags)

> "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."

archiving (& backup) can be an application on top of a synchronization infrastructure.

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

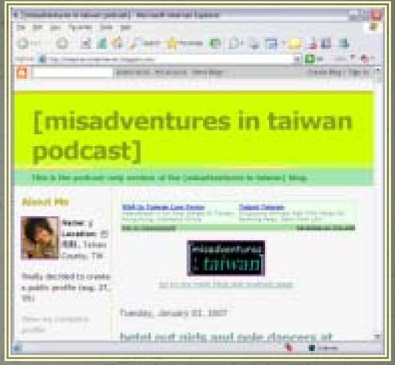
  in other words, where is any digital item? what's its relative worth? what is it? how and when did I get it and when have I accessed it?

- device/repository manager/kb
  - characteristics of each device/repository (e.g. reliability, capacity)
  - access management
  - pings each online element

  in other words, where are my repositories? how do I access them? are they still alive? who else can see them?

- retrieval/re-encounter machinery (e.g. inverted index or metadata db)

podcast only version on blogger

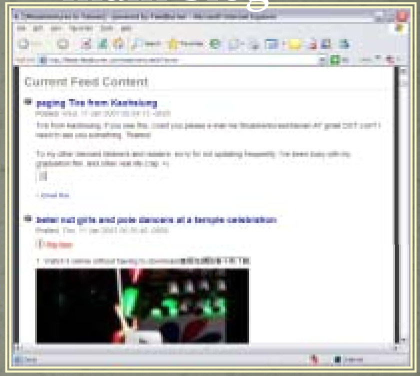


[misadventures in taiwan podcast]

About Me

betel nut girls and pole dancers at

RSS feed from main blog

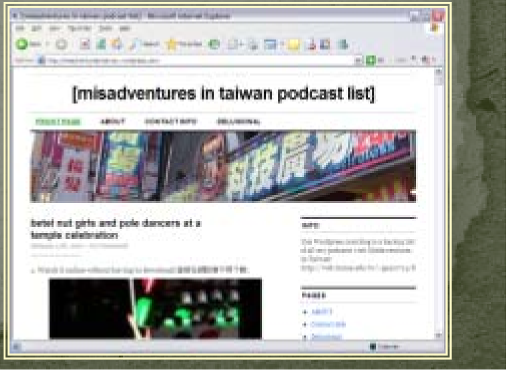


Current Feed Content

paging Tse from Kaohsiung

betel nut girls and pole dancers at a temple celebration

backup blog on wordpress



[misadventures in taiwan podcast list]

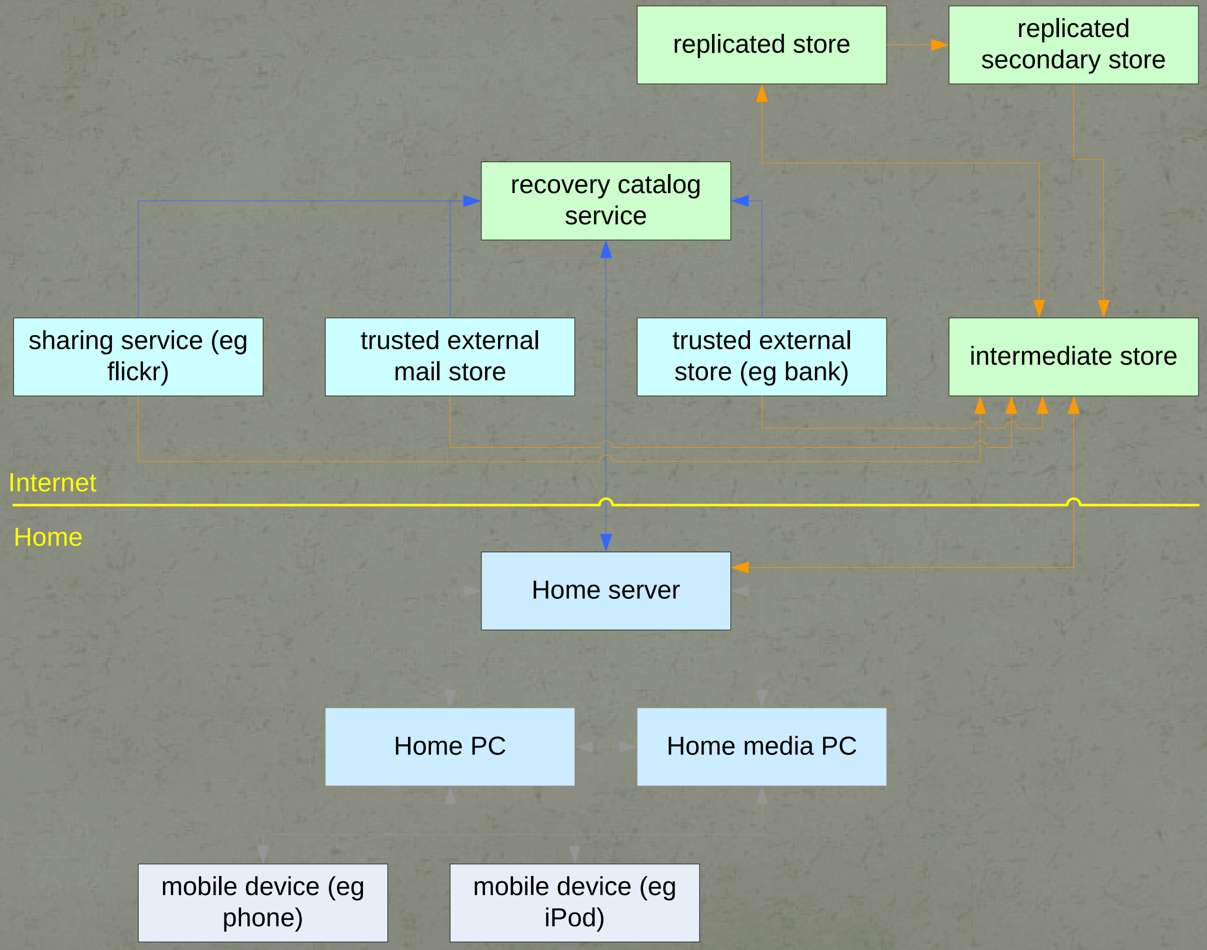
ABOUT

CONTACT INFO

betel nut girls and pole dancers at a temple celebration

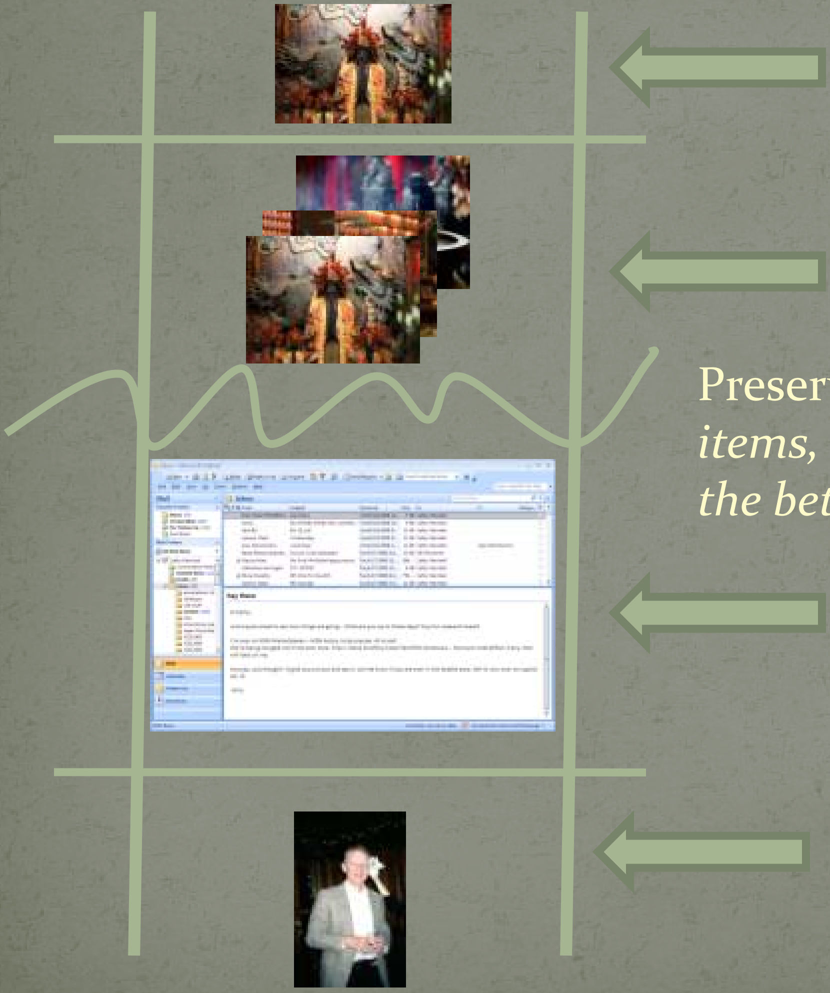
PAGES

an architecture for storing high-value items



what should we keep?

value dictates how many copies we need and where they're stored



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; we don't want to throw it away, but we might not care if we lose it.

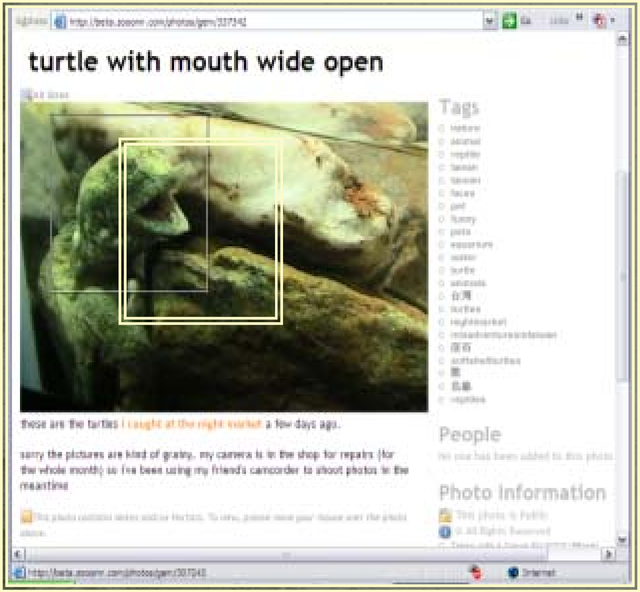
the controversial stuff—deleted? is it emotionally viable to keep everything?

use-based heuristics for assessing value

type	value indicator	example
source	created locally	novel (.doc file)
	p-t-p file	bootlegged music (.mp3 file)
action	edit metadata	name a photo
	view content	play a song
disposition	upload to service	share on Flickr
	remove	drag to trash

how should we maintain it?

curation services and mechanisms



http://beta.zooomr.com/photos/gem/337342

turtle with mouth wide open

these are the turtles i caught at the night market a few days ago.

sorry the pictures are kind of grainy. my camera is in the shop for repairs (for the whole month) so i've been using my friend's camcorder to shoot photos in the meantime

- invisible routinized activities that can be automated as services
  - virus and malware checks (68% of consumers had virus/malware infections in 2005)
  - find files that need canonicalization at deposit (some people save important photos in RAW format)
  - refresh storage media (predictable deterioration)
  - (migration? the jury is out on this one)

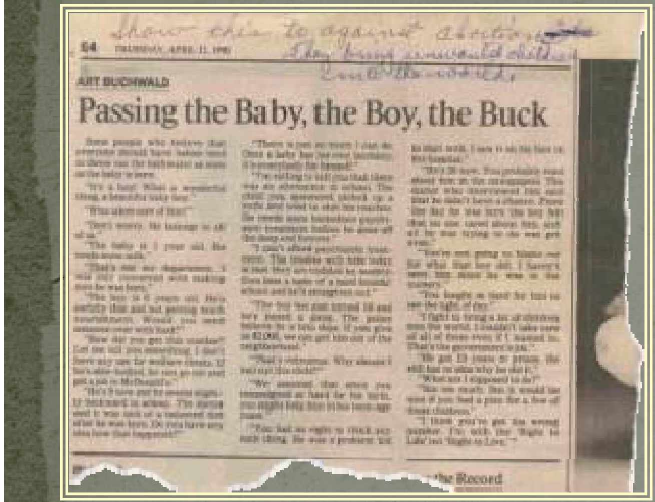
- communal organizing and labeling activities (harnessing the power of social networks)
  - on an individual level, tags, narrative,
  - annotation
  - on an institutional level, format registries

- everything else (stuff requiring human intervention) how will we ever find it again?

(at least) three different cases…

- we’ve forgotten it altogether: re-encounter
- we’ve got some context and a rough idea of what we’re looking for: faceted browsing
- we’ve got a good idea of what we’re looking for: alternative visualizations (and desktop search)

re-encounter: access to *forgotten* stuff



Show this to against abortion
they bring unwanted children
into the world.

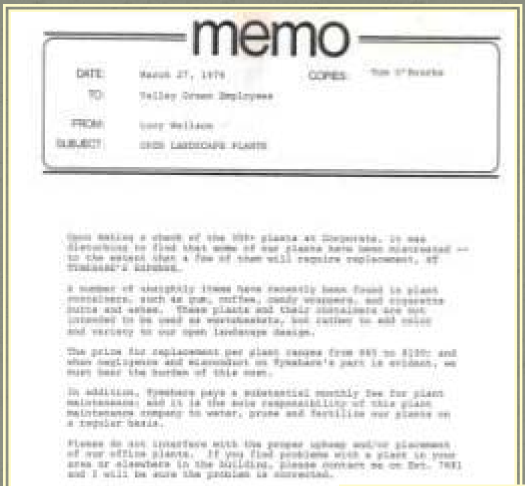
54 THURSDAY, APRIL 11, 1991

ART BUCHWALD

Passing the Baby, the Boy, the Buck

[illegible]

The Record



memo

DATE: March 27, 1979 COPIES: Tom O'Rourke

TO: Valley Green Employees

FROM: Lucy Wallace

SUBJECT: OPEN LANDSCAPE PLANTS

Upon making a check of the 300+ plants at Corporate, it was disturbing to find that some of our plants have been mistreated -- to the extent that a few of them will require replacement, at Tymshare's expense.

A number of unsightly items have recently been found in plant containers, such as gum, coffee, candy wrappers, and cigarette butts and ashes. These plants and their containers are not intended to be used as wastebaskets, but rather to add color and variety to our open landscape design.

The price for replacement per plant ranges from $65 to $150; and when negligence and misconduct on Tymshare's part is evident, we must bear the burden of this cost.

In addition, Tymshare pays a substantial monthly fee for plant maintenance; and it is the sole responsibility of this plant maintenance company to water, prune and fertilize our plants on a regular basis.

Please do not interfere with the proper upkeep and/or placement of our office plants. If you find problems with a plant in your area or elsewhere in the building, please contact me on Ext. 7681 and I will be sure the problem is corrected.

Your anticipated cooperation is appreciated.

Lucy Wallace

[illegible]

TYMSHARE

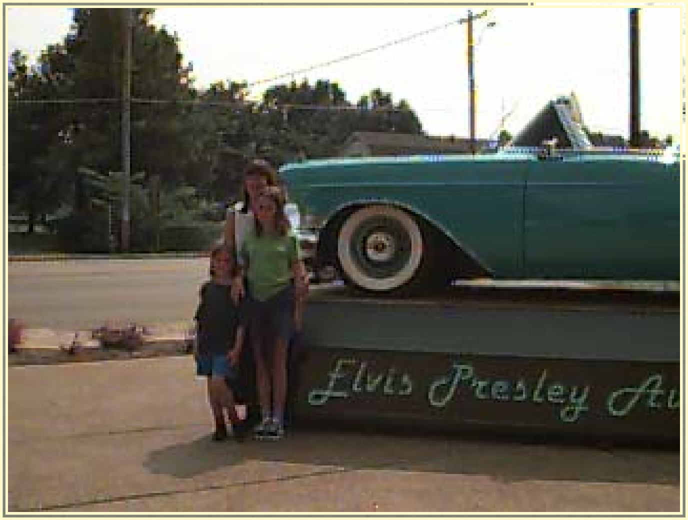
SCHOLASTIC

NEWS RANGER

...ews of President Nixon

...n is the 37th ...nited States. ...dent for four

On January 20, there will be a big parade and party for the President. Maybe you and your family will watch him on TV.



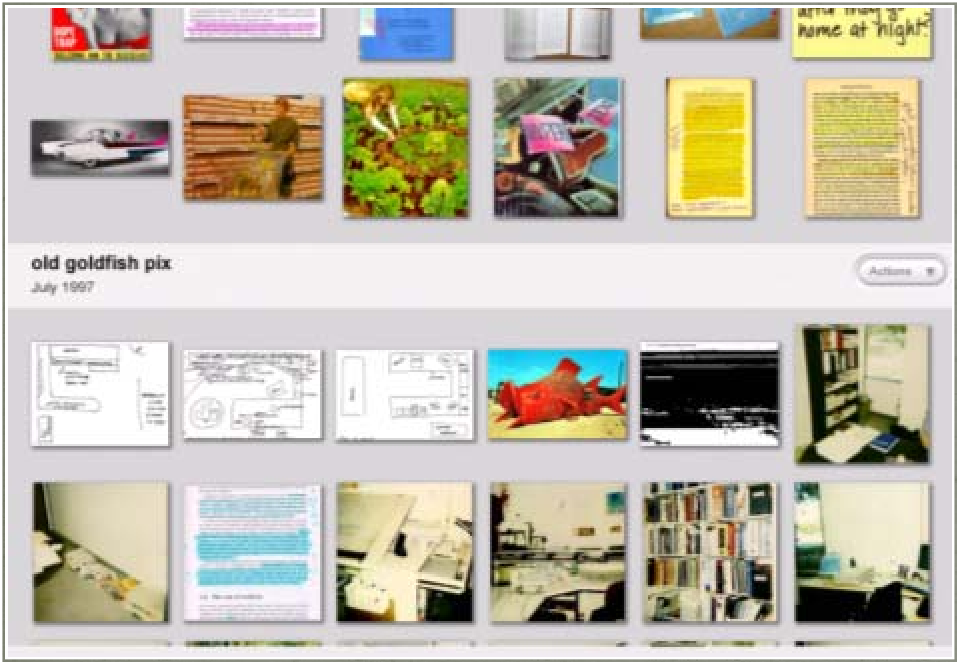
re-encountering

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.

Re-encounter is probably more effective if the item is either in-context (i.e. Implicit Query-based) or high-value (browser-based).

techniques for re-encounter



stable personal geography

- differentiated places

value-based organization

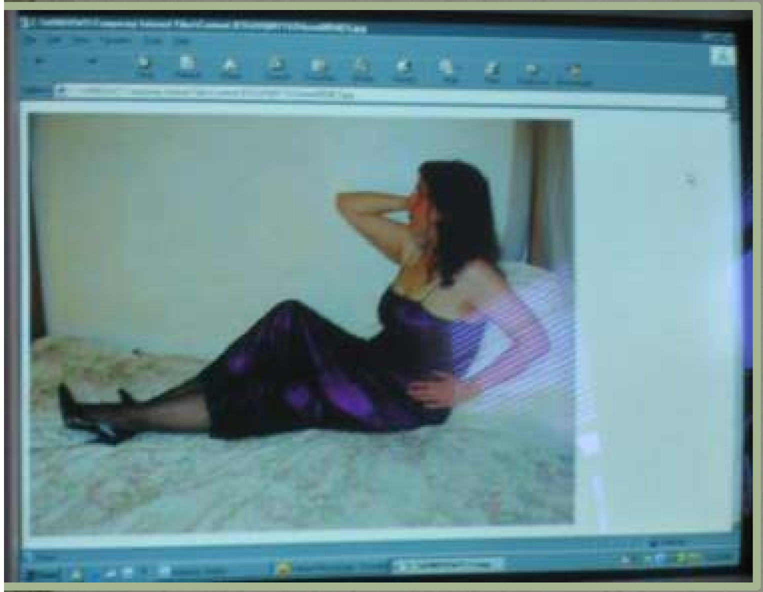
- re-encounter of high-value items

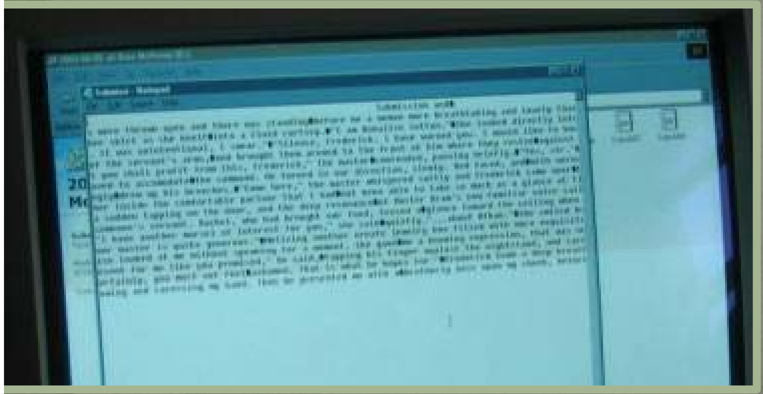
better presentation of item surrogates

- develop good reduced representations of media types other than photos!

(implicit query)

But re-encountering techniques must be approached with care…





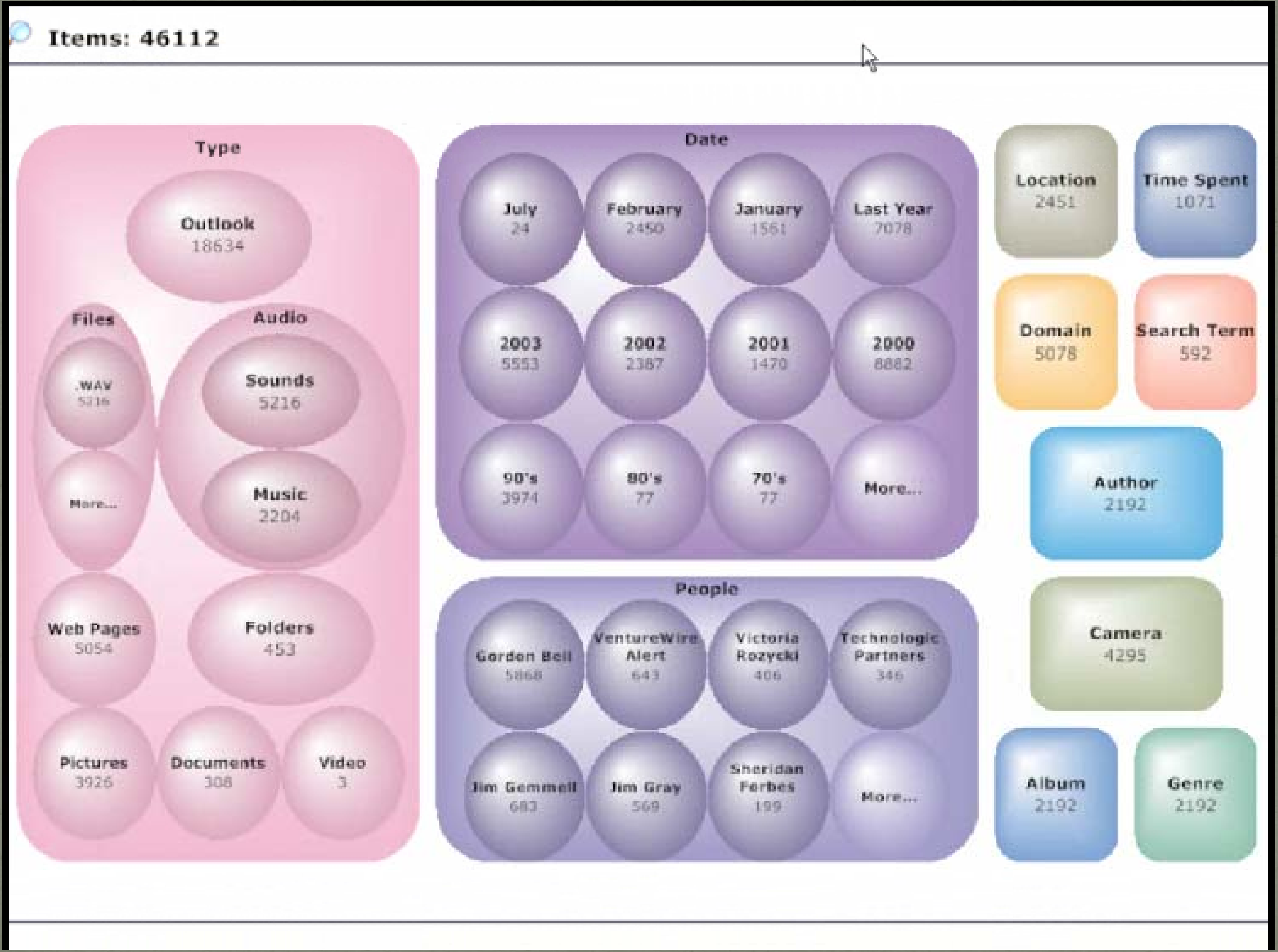
"Oh, it's looking at all the hard disk. … [Clicks on a photo.] Ooops! Sorry! I'm ready to commit suicide."

"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."

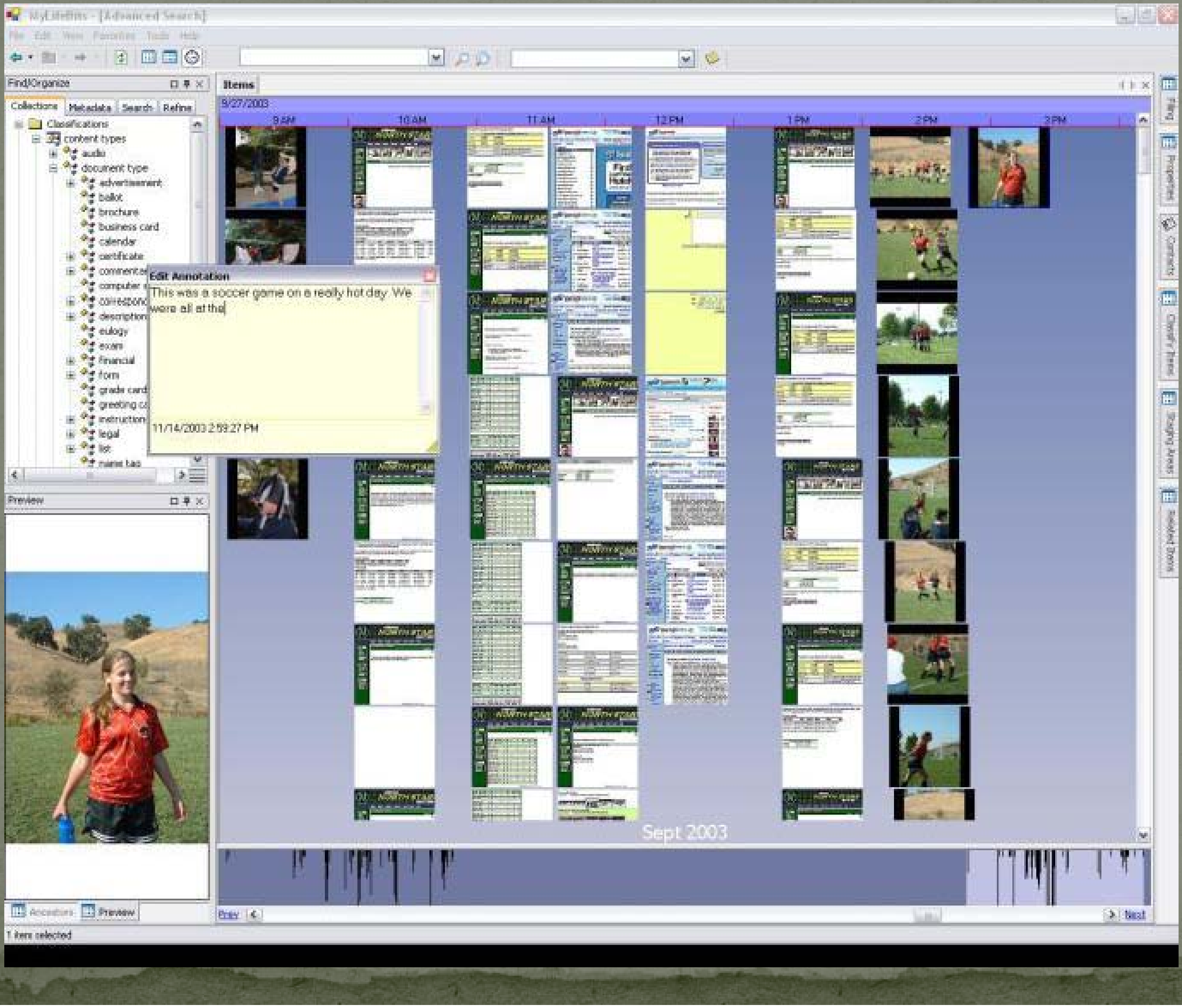
"I have, umm, erotic photos which every man downloads."

"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?'"

faceted browsing (from myLifeBits)



alternative visualizations: e.g. annotated time line (also from myLifeBits)



Edit Annotation

This was a soccer game on a really hot day. We were all at the

11/14/2003 2:59:27 PM

browsing tools should help you select among copies and near copies; they should help you sort out matters of provenance…

whaddya trying to do here, boil the ocean?

this solution should work with other free/for pay solutions

- most archiving software creates yet another storage site that can be cataloged to enhance our approach
  - e.g. free software to create gmail attachments
  - e.g. free software to create S3 backup
  - e.g. for-pay "vault" software

- our approach handles medium- and low-value items
  - can easily be blended with other types of solutions, e.g. vaults for high-value items

- any solution must acknowledge key insights
  - people don't want to keep everything forever
  - nor do they want to decide what to keep and what to throw away
  - nor do they want to pay for storage if they think they have enough copies

miss congeniality (i.e. feel free to write me)

contact info:

cathymar@microsoft.com

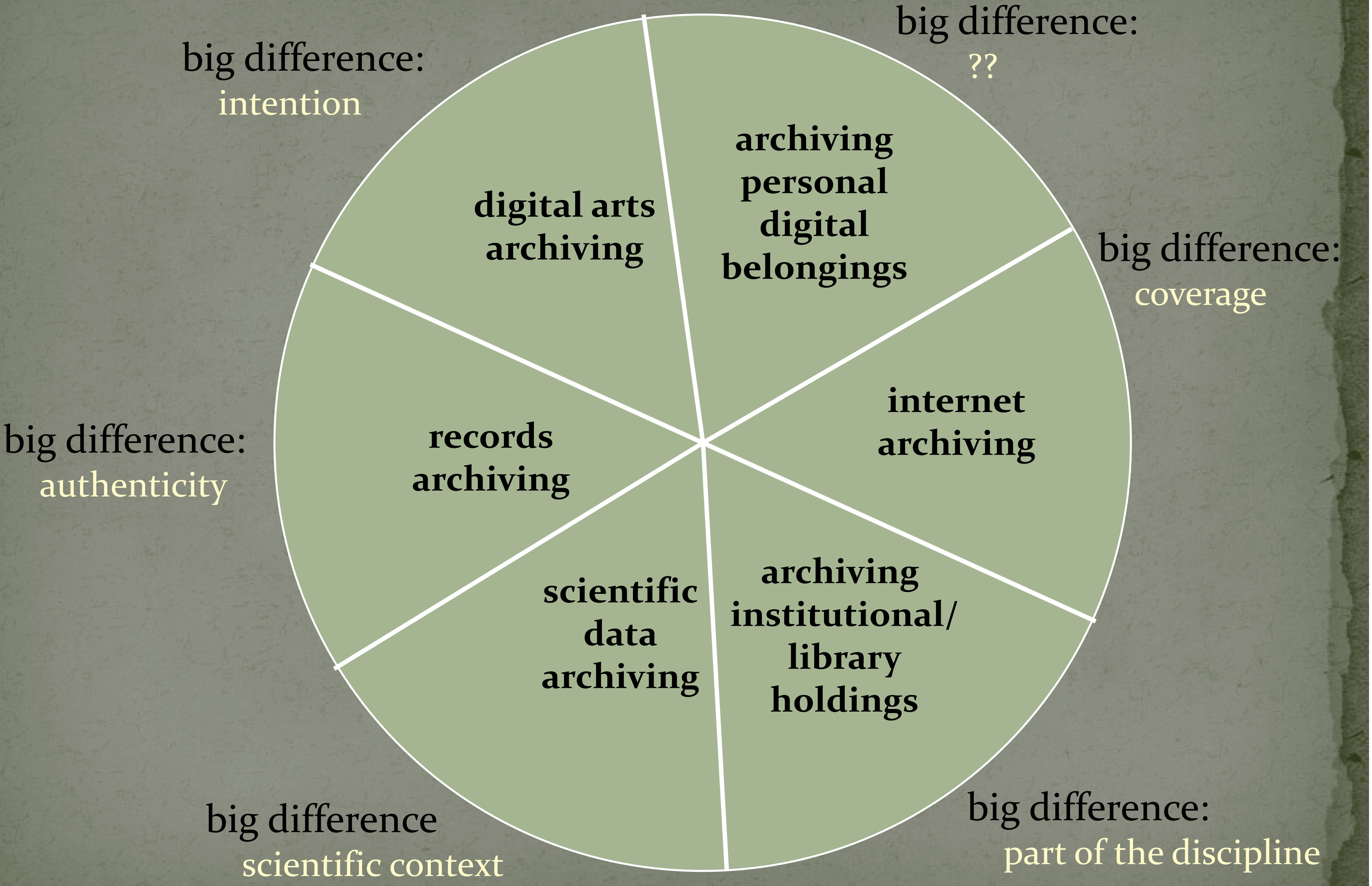
marshall@csdl.tamu.edu

nomiddleinitial@gmail.com

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

http://research.microsoft.com/~cathymar

extra slides



big difference:
intention

digital arts archiving

big difference:
??

archiving personal digital belongings

big difference:
coverage

internet archiving

big difference:
part of the discipline

archiving institutional/ library holdings

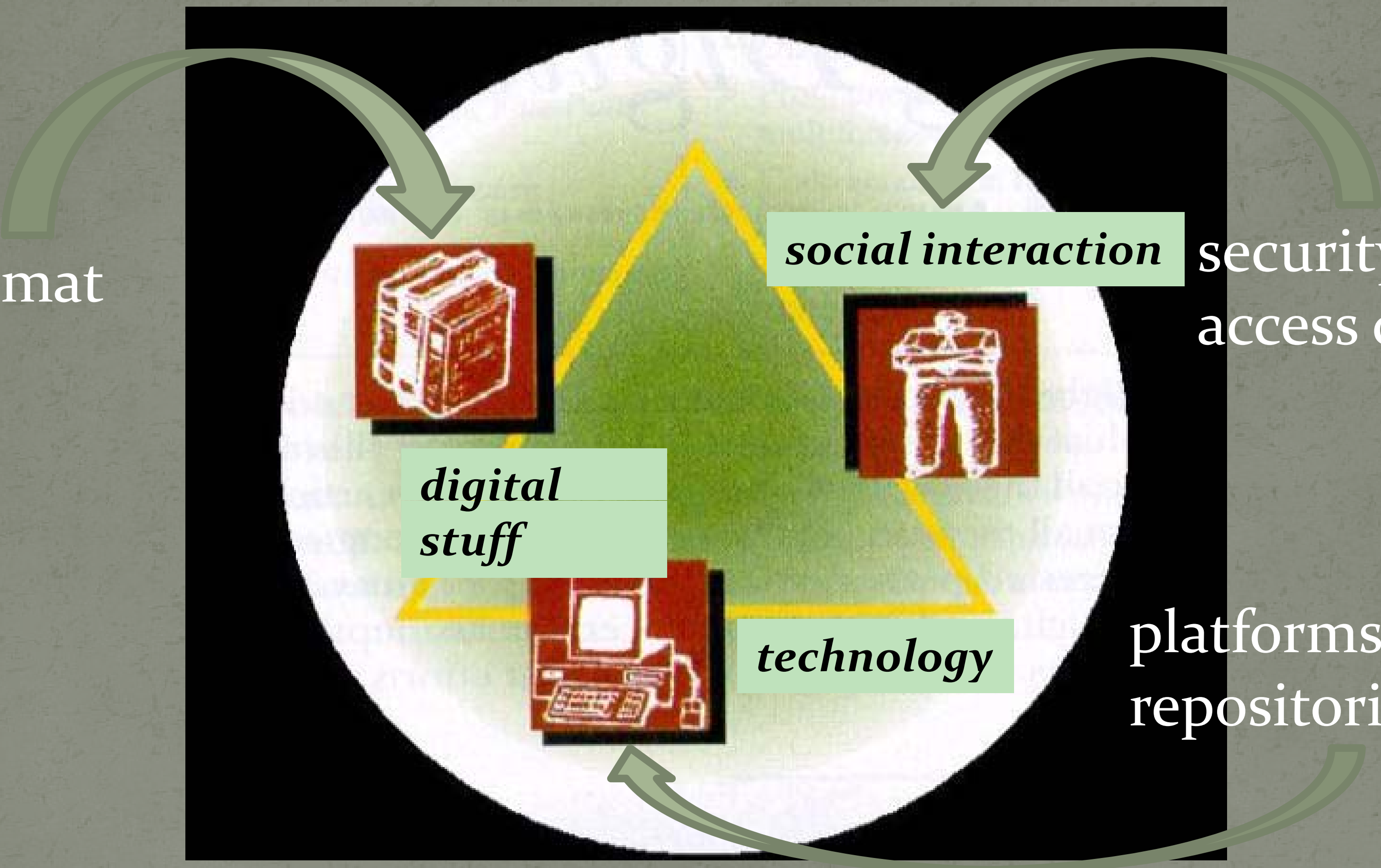
big difference
scientific context

scientific data archiving

big difference:
authenticity

records archiving

How can we find out new solutions to personal digital archiving?



format

digital stuff

social interaction

security and access control

technology

platforms and repositories

by paying attention to all of it:
looking at the whole social/technical sphere...

what do we know?

- personal archiving will have to be a side-effect of normal use
- inferred (and explicit) value will need to help us stratify files to tame the excess stuff
- curation will have to be automated, on-demand, deferred and done through use
- we'll have to provide new modes of searching, browsing, and re-encountering digital stuff