Whose content is it anyway?

Social media, personal data, and the fate of our digital legacy

Cathy Marshall
Microsoft Research, Silicon Valley

TPDL 2012 ANNOTATED VERSION
26 September 2012
first: thank you for coming to this early AM session after last night’s drinking & dancing!
3 mysteries of modern life
Mystery 1:
“Backup? We don’t need no stinkin’ backup!”

i.e. what do we back up these days? How do we decide what to keep?
Ted: I used to [back up my computer]. I haven't been that good with this computer. I got this computer [almost a year ago], which is my first new computer in six years. The other one was a laptop, and I used to back it up to an external drive pretty regularly. And I just haven't migrated that regime to this computer. ... I just haven't done it. The inertia is terrible.
and if you clicked on the flag in the tray on my own screen, you’d see that I’m no exception...
Mystery 1:
could it be the tyranny of all our digital stuff?
maybe it’s all mostly in the cloud.
perhaps we share everything we care about?
(aka “share is the new save”)
Mystery 2:

nobody’s worried about saving social media either
“Pinterest is every single magazine with the best pages ripped out.” –Mary, a manager and closet Pinterest user
Mary: I assume that because [Pinterest] is online, it's backed up. Maybe I'm wrong, but I don't feel like this is threatened...

<the interview continues. we move on to other stuff>

Me: Have you ever had one of these cloud services go away on you, or has everything been relatively stable?

Mary: You're stressing me out! I need to back up my Pinterest. It just dawned on me, what if Pinterest died? Where would my wedding go? ... I don't know why I think Pinterest will live forever.
the next day...

Dear Twitter: Please give me my 1200 missing tweets back, or else I'll make vague empty threats in your general direction. Love, Cathy
11:50 AM Jun 16th via web

an argument breaks out...

@jerepick @gingdottwit @jonathankoren Some of my tweets are ephemeral; some (emphatically) are not! I'd give you examples, but they're gone.
12:01 PM Jun 16th via web in reply to jerepick

a happy ending!

12:03 PM Jun 16th via web
in the ensuing year, I documented at least 3 ways to archive tweets

ccmarshall  cathy marshall
Just when the DSM-V is ready to scrap Narcissistic Personality Disorder, along comes http://twournal.com/ (and I can’t wait to use it!)
13 Dec

THAT camp—a digital humanities meeting. Find out about it.
twitter archive "Twapper keeper"

On own server; can also export them to a csv file
BackupMyTweets

Hue McKemmish, “evidence of me” (1996)

1 billion camera phones in the world

According to @jubois, 35 hours of video uploaded to YouTube per week. Were so many embarrassing personal accidents.

What does it cost to endow a terabyte? Complex cost models like Cartwright—what about democratizing the cost. “Buy a brick” rather…
History
Added On: Jul 13, 2011
Total Items: 5,489
Storage: 10.8 MB
Status: OK
Last Backup: about 4 hours ago
Next Scheduled Backup: Jun 03, 2012 16:00:00

Exports
Exports are zipped and emailed to you. Depending on the request volume and the amount of data, exports can take up to 2 days.

Most Recent Items Backed Up

<table>
<thead>
<tr>
<th>Text</th>
<th>Type</th>
<th>Created</th>
</tr>
</thead>
<tbody>
<tr>
<td>@ccmarshall Next time, Let Us Take Your Bags! Visit...</td>
<td>mention</td>
<td>Jun 02, 2012</td>
</tr>
<tr>
<td>@arnicas @joandimicco What happens when they find out...</td>
<td>status</td>
<td>Jun 01, 2012</td>
</tr>
<tr>
<td>@ccmarshall just ask @joandimicco, she's the expert.</td>
<td>mention</td>
<td>Jun 01, 2012</td>
</tr>
<tr>
<td>Best excuse for dressing inappropriately: &quot;United lost my...</td>
<td>status</td>
<td>Jun 01, 2012</td>
</tr>
<tr>
<td>@arnicas @amandacox Makes one appreciate how much work...</td>
<td>status</td>
<td>Jun 01, 2012</td>
</tr>
</tbody>
</table>

Most Recent Backup History

<table>
<thead>
<tr>
<th>Started At</th>
<th>Ended At</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>May 19, 2012 21:03:38</td>
<td>May 19, 2012 21:04:09</td>
<td>Completed</td>
</tr>
</tbody>
</table>
Mystery 2 addenda:

do cloud storage services mitigate risk or do they make it worse?

Are people okay with losing some of their online presence? How about most of it? How about ALL OF IT?
Mystery 3: what’s the deal with reuse
Picking a nose (from the many online)
aka I’ll never need to take another picture
Kim: I did this blog posting on nose tweaking. Because that was a really big thing. That was the ultimate insult that one man would do to another is tweak [his] nose. That was just the worst thing that you could do to another human being...

That was an early blog post [I wrote]. And I thought, "that's a great nose". I just kind of picked the image off the Internet.

Me: Okay. So that's someone else's nose.
Kim: So I co-opted that picture. It was not [in] Wikicommmons. I was bad. And that picture has gotten more hits than almost anything else I've posted on my blog. It is just amazing. I use a picture of a nose, and it gets picked. Worldwide. ...

Me: Actually I can understand that. I've looked for pictures of noses for myself. For my blog.

Kim: So that picture has gone around the world 20 or 50 million times is my guess. Because I just picked it off somewhere.
Mystery 3: reuse.
how has all this user-contributed information changed the way we work?
does casual reuse have implications for institutional archiving?
~70% of the world's digital content is created by individuals

(UK Telegraph, IDC annual survey, 5/2010)
There were 90 billion photos on Facebook as of January, 2011, growing at 6 billion new photos per month.
10% of all photos EVER taken were shot in 2011.

--Fortune Magazine, 9/24/2012 p.166
Almost 90% cite Facebook and 70% cite Twitter as a primary source for news

--2012 Social Media News survey
You’ve got to look at more than just technology or content...
personal information keeping practices have shifted dramatically over the past decade...

- consumer archiving field study (2005) with Sara Bly & Francoise Brun-Cottan  
  surprise: digital benign neglect

- lost website recovery study (2007) with Michael Nelson and Frank McCown  
  surprise: loss often isn’t due to technology

- social media ownership study (2009-now) with Frank Shipman  
  surprise: coalescing social norms for reuse

- online presence study (right now) with Sian Lindley, Abi Sellen, Richard Banks, Tim Regan  
  surprise: people see social media as transient
then

now
Things they said that surprised me (1)...

- **consumer field study (2005)**
  
  “I mean, if we would’ve had a fire, you just move on.”

  “What? You’ve never heard of Whimit?” Irina says with a note of scorn and disbelief in her voice

- **lost website recovery study (2007)**
  
  “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”
Things they said that surprised me (2)...

- **social media ownership (2009-now)**
  
  "[Reusing photos from the Web] is okay most of the time. The only time I would think it isn't would be when the main focus is of someone you don't know. Like when people email out the People of Walmart photos. Those are taken by people who don't know the person in the photo and posted."

- **online presence study (right now)**
  
  "Well. I just can't think of anything in Facebook that I want to archive."

  "My dog probably uses Facebook more than me."
2005
digital *benign neglect*
and its side effects
“[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.”

[do you ever get rid of digital stuff?]

“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.”
“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]”
the family archivist $\neq$ the family IT support person $\neq$ the family photographer

“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 registry is a battlefield...*
personal digital archiving

==/==

archiving a personal digital collection
i.e. Michael Joyce archiving his own digital stuff

==/==

Gabby Redwine archiving Michael Joyce’s digital stuff
Is accumulation really a problem? Storage is cheap and getting cheaper. Why not just keep EVERYTHING?
There are more and more places to store it!
Store it and forget it...

Microsoft® SkyDrive
Dropbox
iCloud
amazon web services™
GoogleDrive
Who are the digital hoarders?
They are us.
It’s easier to *keep* than to *cull*...

but it’s easier to *lose* than *maintain*.
2007

our personal data

has a life of its own
[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] ccm 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 & audiences.

data safety is a side effect!
distributed copies have lives of their own...

downloaded 387 times

3,869 views, ★★★★★★

45 views, no “likes”

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

Each copy grows
• different author-contributed metadata
• different social metadata
• different collections/organization
People use circular reasoning

People think of the local copy as archival (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 online copies have been organized and given metadata

“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.”
“nearly 70% of [reported data loss in the home] manifested itself in an inability to find information; by comparison ... [only] 8% [of loss was] due to hard drive failure.”

—British Library’s Digital Lives Synthesis
Social curation in the cloud elicits metadata -but- it does not ensure data safety
2010

the ownership & control of all this user-contributed stuff is messy business!
The saga of a giant catfish
The giant catfish is from Texas from the bottom of the Hoover Dam from the Mekong River from Mississippi from Lake Texoma or perhaps...

...it’s from the Johnson Family Photo Album

And it weighs...

187 lbs
140 lbs
424 lbs
187.4 lbs
Or perhaps... 646 lbs
Some reusers were content to simply reduce the resolution and crop the fish...
A surprising number added a watermark as a hedge against further reuse...
Of course those with a better command of Photoshop could do more damage...
... Or make more of a statement: the nuclear accident series

Pêche miraculeuse à Tchernobyl

Достоинства Ядерной Энергетики

Kodolenergija

TE ARI IR MUSU LEPNUMS PAR PADARITO
And then, there were those who fully repurposed our catfish...

from the highly anticipated Ken Burns’ documentary “Fish”
from Lessig, Code version 2.0, 2006
What would we like to know about ownership and reuse?

1) What are the emerging social norms that govern reuse? (Creative Commons is a great idea. Why isn’t it working?)

2) What are peoples’ concerns when archival institutions absorb personal digital content?

3) Can institutions do anything to mitigate the anxiety?
a series of studies using Mechanical Turk
1252 responses in total
1090 unique respondents
~200/media type (except photos)
surveys are long (40-50 questions)
mix of demographic, scenario-based, practice
Likert-scale, multiple choice, open-ended
Open-ended questions reveal the diversity of respondents’ online lives

- respondents are social media-savvy and cut a broad swath.
  - “blue-collar” information workers
  - hoping to be entertained
  - gamers, shoppers, youtubers...
- Almost everybody shares and publishes content
- Profiles count as published content
- A spectrum of privacy attitudes and practices
<table>
<thead>
<tr>
<th>What?</th>
<th>#</th>
<th>Example from survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>status updates</td>
<td>805</td>
<td>I publish picture of my children and some status updates of how my day is going or how it was, but never what my plans are. [ED077]</td>
</tr>
<tr>
<td>photo/vids</td>
<td>728</td>
<td>Since im a part time model, i mostly post my portfolio pictures [PH234]</td>
</tr>
<tr>
<td>new factual content</td>
<td>388</td>
<td>How to: Electrical wiring and Auto Mechanics [PC168]</td>
</tr>
<tr>
<td>republish</td>
<td>282</td>
<td>videos from funny or die or from Youtube - also share videos related to education for nurses (my former career) [ED068];</td>
</tr>
<tr>
<td>social media profiles</td>
<td>233</td>
<td>I share ... personal information such as likes, dislikes, name, location, DOB, occupation, etc. [PH140];</td>
</tr>
<tr>
<td>creative content</td>
<td>206</td>
<td>I am a member of Deviantart and publish my artwork on that site regularly. [VC149]</td>
</tr>
<tr>
<td>none</td>
<td>51</td>
<td>I don't share much but I observe others through facebook. [PC202]</td>
</tr>
<tr>
<td>other</td>
<td>38</td>
<td>Scientific data [RE119]; Code, small scripts or programs. [PC147]; WISH LISTS ON AMAZON [PH156]</td>
</tr>
</tbody>
</table>
I share music, interesting or funny pictures I come across, videos, jokes.

I share movies and videos. Pornography and videogames.

I publish humor stuff on my network of websites.

I share information about deals I find, or interesting articles I come across.
As you’d expect, most respondents feel you can save anything you see online...
How often do you download photos you find online?

- Never: 40
- Rarely: 27
- Sometimes: 113
- Frequently: 62
How about reviews—how often do you save your own reviews locally?

Bar chart showing:
- Media/creative work: 99
- Product: 71
- Service: 33

Bar chart showing:
- Always: 15
- Often: 12
- Sometime: 34
- Rarely: 32
- Never: 109
subject

bystander

implied photographer
Keeping stuff is uncontroversial regardless of media type
Looks pretty similar for reviews, except...
Saving’s one thing. But what about reuse?
“Someone tagged a picture of me on facebook and I saved the picture and put it into my album because it was a nice picture.”

“my friend took a picture of my son and her son. I reposted so everyone can view from my family.”

“Earlier this week I downloaded a picture of a dog wearing a party hat for a story I was doing on my pet blog about an event coming up. It was a photo included in a press release by the store holding the event.”

“I couldn't find a photo I'd taken on a trip, which I wanted to use in a Facebook album, so I found a photo of the same landmark on someone's blog and republished it in my album.”
A complex ethical and moral terrain

- **law**
  
  "the web is public domain. if you post a picture (with or without a copyright watermark), it's going to be seen and recycled by others. ..."  [PH153]

- **creation context**
  
  "...if it is picture someone takes of a flower, or a sunset, and you find it pretty and want to save it, then that's fine..."  [PH227]

- **reuse context**
  
  "...If it's on a website that makes money but it's not money coming from your photo but your photo makes the website look better then I think that would be okay..."  [PH207]

- **technology**
  
  "I only post photos that feature the 'share' button for Facebook built in to the blog."  [PH140]

Reuse in practice may be mediated by need, the nature of the material, history (e.g. how much it has already ‘been around’)
Factors that influence social norms re: reuse

- Personal experience
  Practical examples v. hyperbole
- Media type
  photos v. tweets v. videos v. reviews, etc.
- Perception of misuse or harm done
  redundancy
  malicious reuse (e.g. fraud, correctness*, ‘making fun of’)
  children as protected class
  commercial value
- Mis/understanding of applicable law
  online=public domain
  permission
2012

wait! People see their social media as transient!

Not only is it transient; they may destroy it themselves.
some annotated excerpts of a participant's map
Kim: [drawing map] Then there's Facebook. I'll make it small...Huh. Well. I just can't think of anything in Facebook that I want to archive.

Mary: If my facebook died, all those photos live somewhere else and I don't particularly care about the comments that much.

Ted: My dog probably uses Facebook more than me...I don't really care what people are doing on [Facebook]. I go there out of social obligation.
me: **Do you ever save anything from Facebook?**

Mary: You mean photos, or do I ever take screenshots of conversations?

me: **Any of those things.**

Mary: No. I've never found a need because I can always scroll back in the history. And the photos are in albums. So, no.

me: **Do you back it up?**

Mary: No. No. Yes. Yes. We have a Lamborghini hard drive at home.

me: **How do you back up [Facebook]?**

Mary: That's an excellent question. I don't back up Facebook.
me: Is Facebook your primary thing?

Lynn: Unfortunately, yeah. I’m on it more than I want to be. [laughs]...so – my friends are in LA and Vienna and the east coast, so it’s a nice easy way to keep track...

me: Do you ever want to save any of this stuff?

Lynn: It’s all saved. On here.

me: In the sense that, you don’t expect Facebook to go down in flames anytime?

Lynn: No. And if they do—y’know—no, there’s nothing on here that I’d like to save.
But is that social media metadata really so unimportant?

Kim: It's funny but how, when you go back, are you going to remember what it was? Like I already look at stuff and I can't tell if it was Veronica or Abby's baby pictures. I didn't write it down. It's really pathetic.

It's like, this looks like Venice. I think it was Venice.
the case for institutional archiving

why we might expect people to need some help sorting this out...
People have shown considerable interest in outreach efforts, classes, products...

Solve your photo problem! Taking photos has never been easier. But what do we do with all of our digital images? And who gets to see them?

At Heritage Makers, we help people preserve, share and celebrate their photos with easy-to-use online software. With Photo Studio, you can rescue the photos on your hard drive, create albums and slideshows, and share them with friends and family. Whether you choose professionally bound books, a personalized gorgeous, wrapped canvas print, or digital keepsakes with Heritage Makers, you are in control of the outcome.

Backup and Restore for Social Media

Simple, automated backups of your personal, Facebook, Gmail, Twitter and other social media & web apps accounts. Perfect for personal users, small businesses and social media professionals.
"If `ifs' and `ans' were pots and pans ... "

- **disaggregation of skills**
  The family archivist \(\neq\) family IT person

- **trends in personal data management**
  Social media + cloud stores + multiple devices

- **overwhelming power of benign neglect**
  Rapid accumulation
  Digital hoarding
  Ambiguous value of both local and social stuff
  Creation is more rewarding than stewardship
the physical/digital break
first digital photos—rules still apply
photos from my phone—artistic and emotional anarchy—rules evaporate
where does all of this leave us? Should institutions snap up social media? Will Memento-like services fill in the gaps?
from a recent-ish Library of Congress video

kid 1: “They should just save Facebook. That is our generation’s scrapbook, yearbook, Guinness World Record...”

kid 2: “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.”
But what happens when institutions just snap up a feed?

“This is what makes Twitter’s “gift” troubling. It assumes that all content shared publicly is truly public and for posterity. ... [Consider this scenario:] Bob wants to be practically obscure – private in public – without going to all the trouble of setting up complicated privacy controls. So what happens, two years from now, when Bob accidentally discloses his handle in the wrong context, and he needs to remove some Tweets?”

-Fred Stutzman, blog post on 4/14/2010
@zittrain "if what you get online is free, you're not the customer - you're the product"
It’s not storing social media that gets people worked up.

*It's the potential for access and reuse*

But access and reuse (for a variety of purposes) may be the best justification for institutional archiving.
attitudes toward institutional archiving

why we should pay attention to them ...
Using Library of Congress as a proxy for other public or non-profit institutions
Our surveys each included a LoC scenario per media type. (1252 respondents)

**Survey 1:** The Library of Congress is acquiring the public Twitter feed, dating back to the site’s origins. They are planning to provide access to the archive. (173 responses)

**Survey 2:** ditto Flickr. (242 responses)

**Survey 3:** ditto Amazon book reviews. (203 responses)

**Survey 4:** ditto iTunesU. (209 responses)

**Survey 5:** ditto iTunes podcasts. (225 responses)

**Survey 6:** ditto YouTube. (200 responses)
3 contrasting access conditions

- Strongly Agree
- Agree
- Agree somewhat
- Undecided
- Disagree somewhat
- Disagree
- Strongly disagree

Only researchers can access the archive now.

Everyone can access the archive now.

Everyone can access the archive in 50 years.
Should the Library of Congress be able to archive any social media that’s available today on the public Internet?

What should be excluded and why?

http://brutal-knitting.tumblr.com/
## 4 perspectives

<table>
<thead>
<tr>
<th>Creator</th>
<th>Content/media</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <em>permission</em></td>
<td>• <em>inherent value</em></td>
</tr>
<tr>
<td>• <em>credit</em></td>
<td>• <em>veracity</em></td>
</tr>
<tr>
<td>• <em>privacy</em></td>
<td>• <em>harm/bias/malice</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Technology</th>
<th>Legal/social</th>
</tr>
</thead>
<tbody>
<tr>
<td>• <em>authorization</em></td>
<td>• <em>public record</em></td>
</tr>
<tr>
<td>• <em>settings</em></td>
<td>• <em>public domain</em></td>
</tr>
<tr>
<td></td>
<td>• <em>social good</em></td>
</tr>
<tr>
<td>Permission/Privacy</td>
<td>Value/Veracity/Harm</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>“...Everything created deserves to have proper credit by its creator, and if the creator doesn't want it archived that should be an available choice”</td>
<td>“...No one cares what status updates someone in Colorado writes about the sandwich they ate for lunch.”</td>
</tr>
<tr>
<td>“Facebook pages, Twitter accounts, and any other social media ... It is that person's private account and should not be messed with.”</td>
<td>“... I think the Library of Congress is more useful by keeping information that has reasonably been researched to know is true, as some of these things out there are not real.”</td>
</tr>
<tr>
<td>“…Videos are a lot more personal than anything written on paper so they should be treated more cautiously.”</td>
<td>“I think social media that contains racial bias or any kind of prejudiced based content shouldn't be archived.”</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Existing mechanisms</th>
<th>Law/Social Good</th>
</tr>
</thead>
<tbody>
<tr>
<td>“profiles indicate the user's express wish not to have their information accessible to everyone.”</td>
<td>“Once something is on the Web, it belongs to the Web users.”</td>
</tr>
</tbody>
</table>

“We are creating culture and history. No matter how some people feel about a certain subject or genre, nothing should be excluded.”
So what do we make of all of this?

Ownership boundaries are fuzzy

Social norms around reuse are evolving

Content type and genre matters!
- How personal is it perceived to be?
- How familiar is it?
- How much experience do people have with it in their everyday lives?

Collection-building and access need to be teased apart—people don’t do make the distinction on their own.
“This is just too hard of a problem”

precautions themselves must be approached with caution

- c.f. The Preserving Virtual Worlds project efforts to obtain permission were met with some hostility and an at-best permission rate of 10%
Public institutions have an important role to play

Stewardship

“The Library of Congress should really keep at least a sample of everything. We are creating culture and history. No matter how some people feel about a certain subject or genre, nothing should be excluded.”

Set peoples’ expectations

- Collection boundaries
- Access boundaries
- Identity and attribution
- Reuse scope
2005 study: Sara Bly & Francoise Brun-Cottan

2007 study: Michael Nelson & Frank McCown (ODU)

2009 + study: Frank Shipman (TAMU)

2012 study: Siân Lindley, Abi Sellen, Richard Banks, Tim Regan (MSR, Cambridge)
contact info:

cathymar@microsoft.com

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

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

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