The Traces We Choose: Selection Strategies by, for and about Individuals

Cal Lee

School of Information and Library Science

University of North Carolina, Chapel Hill

Saving Our Present for the Future: Personal Archiving 2010

Internet Archive

February 16, 2010

San Francisco, CA

UNC
SCHOOL OF INFORMATION
AND LIBRARY SCIENCE

Human activities leave traces

Doors left open

Footprints in the sand

Receipts

Browser cache contents

Temp files

Headers of IP packets

Pixels on the screen of a GPS device

Voice mail messages

Flight prices on an airline web site

All of the traces can convey information

The vast majority of traces only play a role in their immediate context & then disappear

But sometimes we want the traces to stick around for a while.

Any time we purposely increase the chance of use across contexts, we’re engaging in **selection**.

As individuals, we can select by:

- Creating **richer traces** of some moments (e.g. high-res photo of event)
- Making **extra copies** in multiple places
- Storing information in **multiple ways** (e.g. online services, formats, systems)
- **Exporting** information, to reduce the risk of lock-in
- **Sharing** with other people, so they have copies, too

“Collectors” can select by:

- **Capturing** information from dynamic environments (e.g. web archiving)
- Developing and maintaining long-term **collections**
- **Value-added** actions on parts of collections (e.g. search & analysis capabilities, technical support, sophisticated transformations)

Resources are limited, and meaning is expensive.

All transfer of meaning across contexts has **costs**.

Richness and Internal
Complexity of Meaning
Being Transferred

× Degree of Difference
between Contexts

Total Cost to Transfer Meaning
Across Contexts

To advance the curation of personal digital archives, we can:

Get, Grab, Guide

Get
(On Removable Media)

Many collecting institutions have received entire computers from donors.

More common is the “disk in a box” – removable media obtained along with analog collections.

The media will inevitably become unreadable (if they haven’t already).

Librarians/archivists must extract useful information from the media, while avoiding accidental alteration of data or metadata.

Examples of Selection Decisions:

- Low-level Data on Disk – create bit-level image of disk or copy files through filesystem?
- MS Word Document – retain “hidden” data or only what ones assumes to be text author intended?
- MS Outlook .pst File – retain whole file (includes calendar, drafts, deleted messages, address book) or extract sent/received messages and attachments?

Grab
(From the Internet)

“…your identity on the computer is the sum of your distributed presence.”

Turkle, Sherry. *Life on the Screen: Identity in the Age of the Internet*. New York: Touchstone Books, 1997. p.13

To be available for future use, Web content must be:

- continuously maintained by host
- preserved by a distributed set of individuals involved
- harvested by someone with an interest in collection building (amateur enthusiast, interested scholar, archivists, librarian)

Cloud Computing: A Frank Definition

“...a vital, distinct part of what you do and what you’re about or what you consider important to you is on other machines that you don’t run, don’t control, don’t buy, don’t administrate, and don’t really understand.”

Scott, Jason. “Fuck the Cloud.” ASCII. January 16, 2009. http://ascii.textfiles.com/archives/1717

There are many risk factors associated with reliance on web service providers for persistent access to personal materials:

Expiration of service

After a period of user inactivity (e.g. no login or purchase), data deletion is triggered (common in free photo sharing sites)

Changes in service offerings

Companies going out of business

Take-down based on complaints from other parties, e.g. claims of

- Obscenity
- Intellectual property
- Right to publicity
- National security
- Confidentiality

Mergers of sites/companies, resulting in major displacement or loss of content

Drive failure with insufficient backup

Purposeful destruction by malicious attackers

There have been widely-publicized cases of **all** of these types of loss in recent years.

Five Ways to Collect Web Content

- Ask the provider
- See if someone else has it
- Follow links
- Pulled via queries
- Pushed via queries

All Five Ways Involve Selection

- Sources to engage
- How often
- Collecting parameters
- How much effort to invest in fixing specific problem cases

Asking the Provider

- Requires a lot of cooperation
- Can yield information not directly accessible through other means
- Can get data directly from the source (e.g. whole database, high-res images) rather than what’s served through the Web

Seeing if Someone Else Has It (Warrick* Model)

- Has it been:
  - – Cached by a search service (e.g. Google, Yahoo)?
  - – Harvested by Internet Archive?
  - – Collected by a peer institution?
- If so, you could get a copy from them

*Warrick – Recovery Your Lost Website. http://warrick.cs.odu.edu/

Following Links

- Start with seed URLs, then recursively follow them – possibly feeding new URLs back into seed list
- Used by search engine bots and many web crawlers

Pulling via Queries

- Submitting queries to known sources
- Main selection factors:
  - Sources to query
  - How often to query
  - Query terms to use
  - Query parameters (e.g. date, source, number of in-links)
  - Threshold values for parameters (e.g. only get top 100)
- Example: VidArch project posed queries daily to YouTube, collecting top 100 results for each

Pushing via Queries (Subscription)

- Tapping into alert services
- Examples du jour: RSS, Atom, Twitter
- Particularly good for communication forms (e.g. blogs) that are “post-centric” rather than “page-centric”*

*Attributed to Meg Hourihan in: Gillmor, Dan. *We the Media: Grassroots Journalism by the People, for the People*. 1st ed. Sebastopol, CA: O'Reilly, 2004.

We still have a lot to learn about "Web Presence Identification"* for collecting personal archives.

*Bekkerman, Ron, and Andrew McCallum. "Disambiguating Web Appearances of People in a Social Network." In *Proceedings of the 14th International Conference on World Wide Web, WWW 2005: Chiba, Japan, May 10-14, 2005*, edited by Allan Ellis and Tatsuya Hagino, 463-70. New York, NY: ACM Press, 2005.

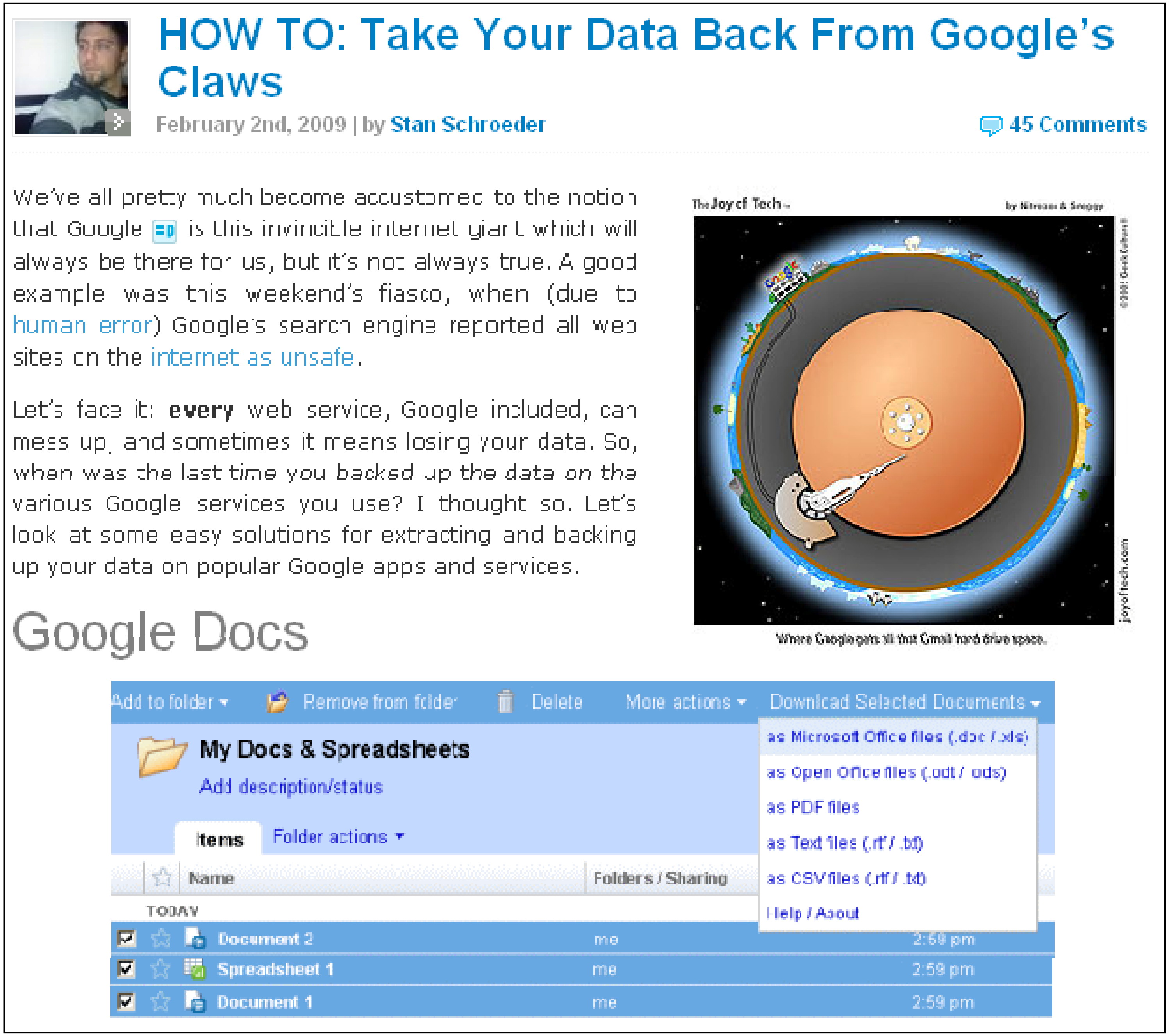
Finally, we can…

Guide

Professional curators will only ever have responsibility for a tiny sliver of documentary traces of individuals.

Helping them to curate their own materials can be as important as taking custody of them.

So maybe we need more things like these:



HOW TO: Take Your Data Back From Google's Claws

February 2nd, 2009 | by Stan Schroeder

45 Comments

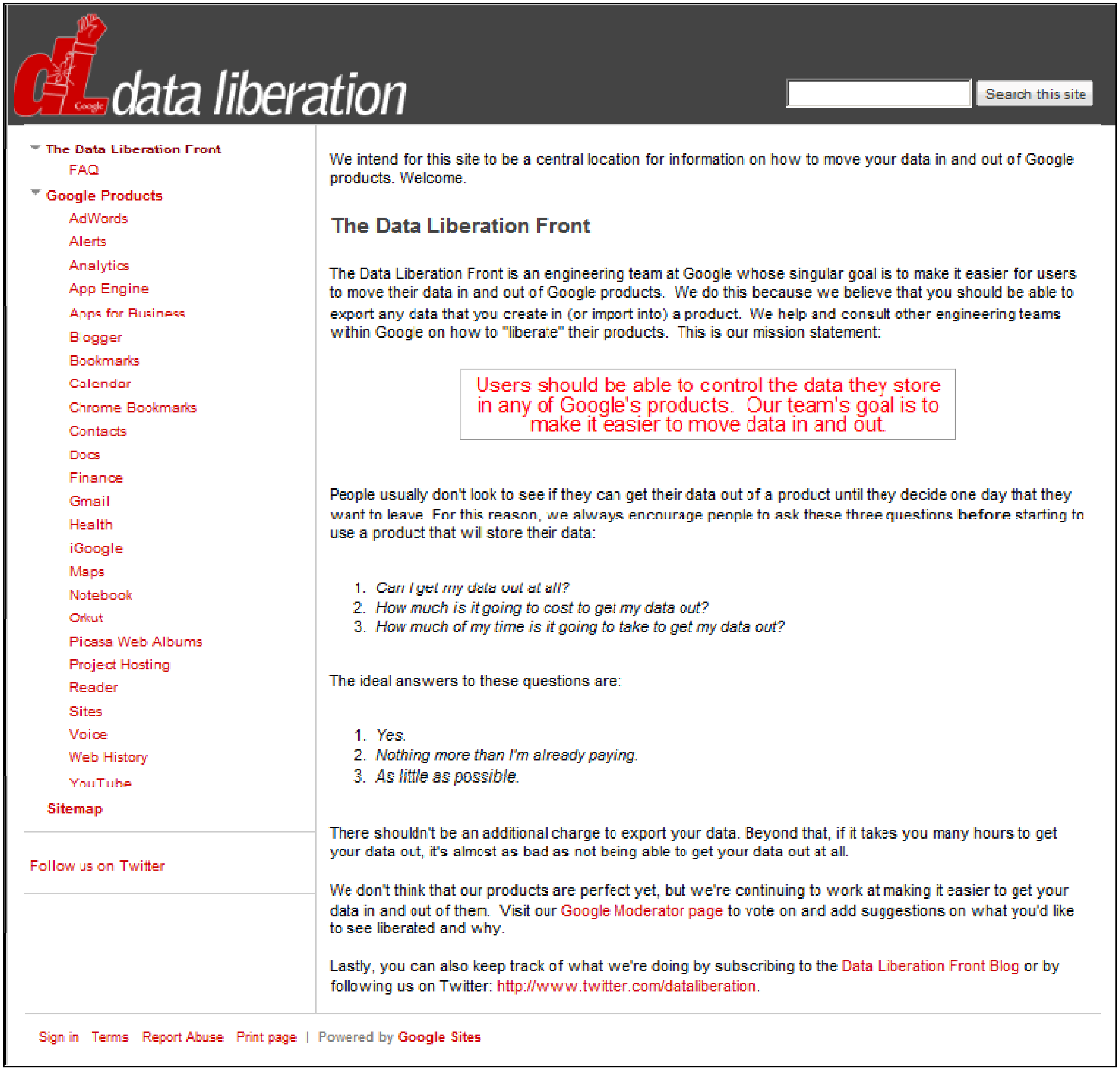
We've all pretty much become accustomed to the notion that Google is this invincible internet giant which will always be there for us, but it's not always true. A good example was this weekend's fiasco, when (due to human error) Google's search engine reported all web sites on the internet as unsafe.

Let's face it: **every** web service, Google included, can mess up, and sometimes it means losing your data. So, when was the last time you *backed up the data on the* various Google services you use? I thought so. Let's look at some easy solutions for extracting and backing up your data on popular Google apps and services.

Google Docs

	Name	Folders / Sharing	
☑	Document 2	me	2:59 pm
☑	Spreadsheet 1	me	2:59 pm
☑	Document 1	me	2:59 pm

http://mashable.com/2009/02/02/google-backup/



We intend for this site to be a central location for information on how to move your data in and out of Google products. Welcome.

The Data Liberation Front

The Data Liberation Front is an engineering team at Google whose singular goal is to make it easier for users to move their data in and out of Google products. We do this because we believe that you should be able to export any data that you create in (or import into) a product. We help and consult other engineering teams within Google on how to "liberate" their products. This is our mission statement:

> Users should be able to control the data they store in any of Google's products. Our team's goal is to make it easier to move data in and out.

People usually don't look to see if they can get their data out of a product until they decide one day that they want to leave. For this reason, we always encourage people to ask these three questions **before** starting to use a product that will store their data:

1. *Can I get my data out at all?*
2. *How much is it going to cost to get my data out?*
3. *How much of my time is it going to take to get my data out?*

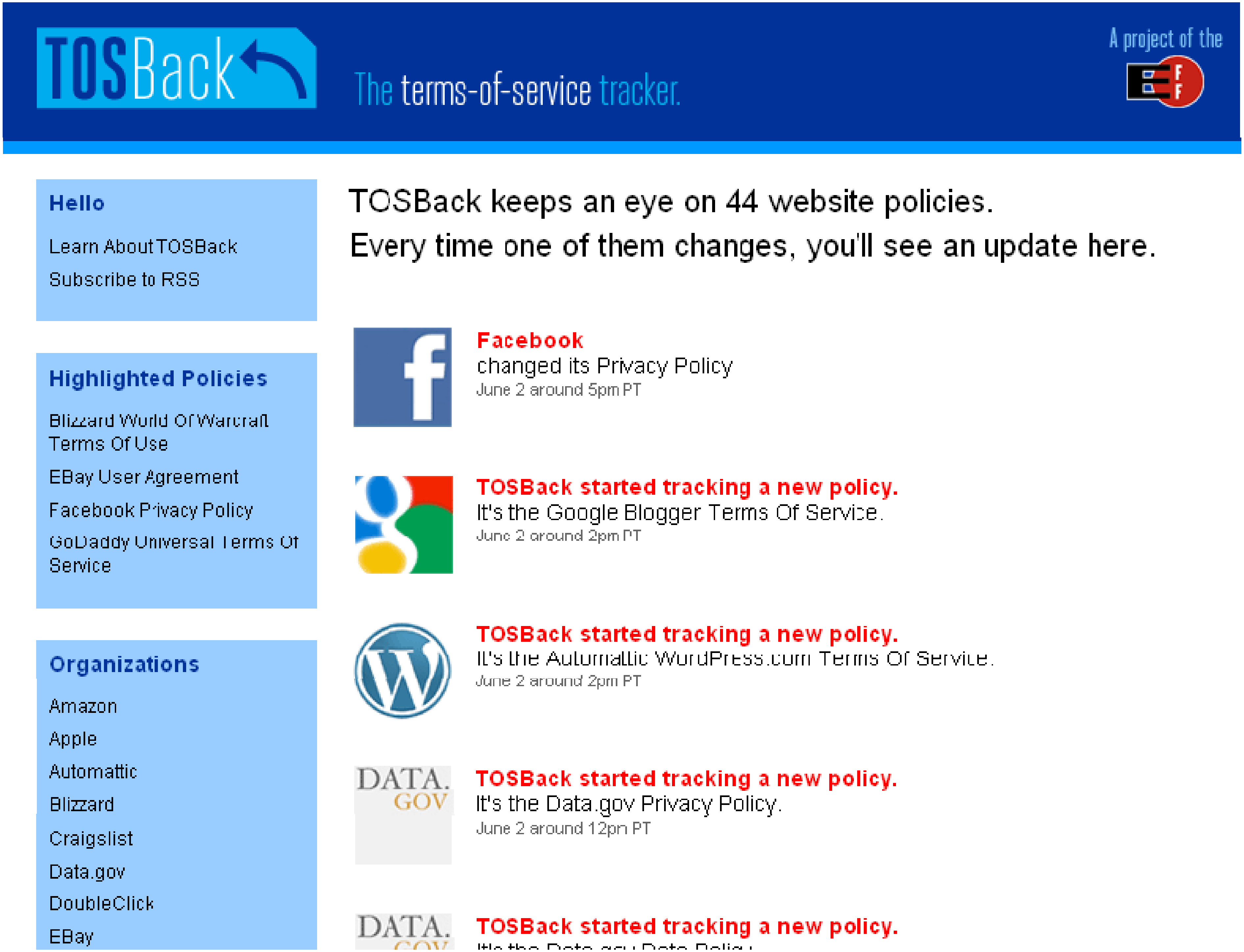
The ideal answers to these questions are:

1. *Yes.*
2. *Nothing more than I'm already paying.*
3. *As little as possible.*

There shouldn't be an additional charge to export your data. Beyond that, if it takes you many hours to get your data out, it's almost as bad as not being able to get your data out at all.

We don't think that our products are perfect yet, but we're continuing to work at making it easier to get your data in and out of them. Visit our Google Moderator page to vote on and add suggestions on what you'd like to see liberated and why.

Lastly, you can also keep track of what we're doing by subscribing to the Data Liberation Front Blog or by following us on Twitter: http://www.twitter.com/dataliberation.



TOSBack The terms-of-service tracker.

A project of the EFF

Hello

Learn About TOSBack

Subscribe to RSS

Highlighted Policies

Blizzard World Of Warcraft Terms Of Use

EBay User Agreement

Facebook Privacy Policy

GoDaddy Universal Terms Of Service

Organizations

Amazon

Apple

Automattic

Blizzard

Craigslist

Data.gov

DoubleClick

EBay

TOSBack keeps an eye on 44 website policies.

Every time one of them changes, you'll see an update here.

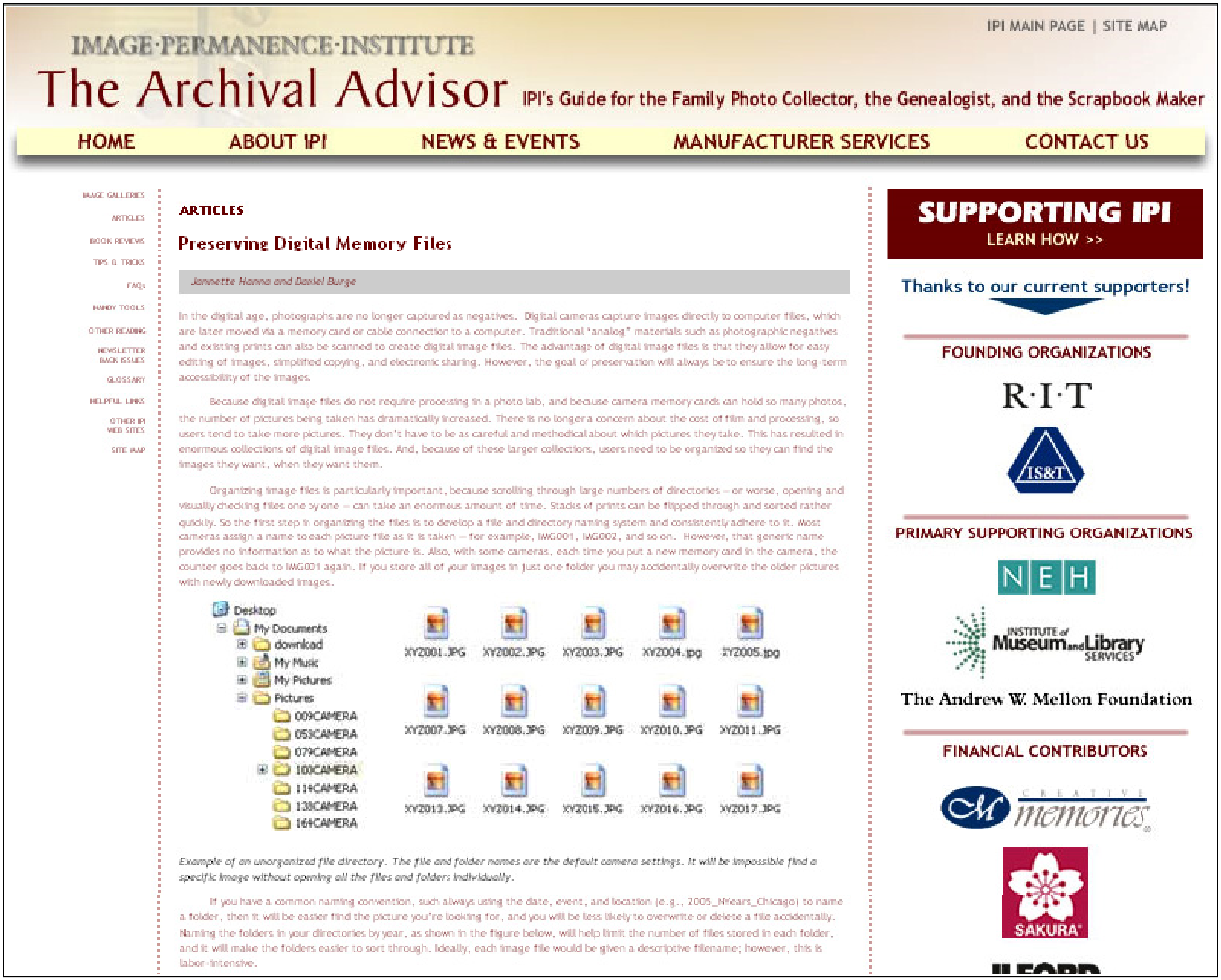
Facebook
changed its Privacy Policy
June 2 around 5pm PT

TOSBack started tracking a new policy.
It's the Google Blogger Terms Of Service.
June 2 around 2pm PT

TOSBack started tracking a new policy.
It's the Automattic WordPress.com Terms Of Service.
June 2 around 2pm PT

TOSBack started tracking a new policy.
It's the Data.gov Privacy Policy.
June 2 around 12pm PT

TOSBack started tracking a new policy.
It's the Data.gov Data Policy



ARTICLES

Preserving Digital Memory Files

Jannette Hanna and Daniel Burge

In the digital age, photographs are no longer captured as negatives. Digital cameras capture images directly to computer files, which are later moved via a memory card or cable connection to a computer. Traditional "analog" materials such as photographic negatives and existing prints can also be scanned to create digital image files. The advantage of digital image files is that they allow for easy editing of images, simplified copying, and electronic sharing. However, the goal of preservation will always be to ensure the long-term accessibility of the images.

Because digital image files do not require processing in a photo lab, and because camera memory cards can hold so many photos, the number of pictures being taken has dramatically increased. There is no longer a concern about the cost of film and processing, so users tend to take more pictures. They don't have to be as careful and methodical about which pictures they take. This has resulted in enormous collections of digital image files. And, because of these larger collections, users need to be organized so they can find the images they want, when they want them.

Organizing image files is particularly important, because scrolling through large numbers of directories – or worse, opening and visually checking files one by one – can take an enormous amount of time. Stacks of prints can be flipped through and sorted rather quickly. So the first step in organizing the files is to develop a file and directory naming system and consistently adhere to it. Most cameras assign a name to each picture file as it is taken – for example, IMG001, IMG002, and so on. However, that generic name provides no information as to what the picture is. Also, with some cameras, each time you put a new memory card in the camera, the counter goes back to IMG001 again. If you store all of your images in just one folder you may accidentally overwrite the older pictures with newly downloaded images.

Example of an unorganized file directory. The file and folder names are the default camera settings. It will be impossible find a specific image without opening all the files and folders individually.

If you have a common naming convention, such always using the date, event, and location (e.g., 2005_NYears_Chicago) to name a folder, then it will be easier find the picture you're looking for, and you will be less likely to overwrite or delete a file accidentally. Naming the folders in your directories by year, as shown in the figure below, will help limit the number of files stored in each folder, and it will make the folders easier to sort through. Ideally, each image file would be given a descriptive filename; however, this is labor-intensive.

SUPPORTING IPI
LEARN HOW >>
Thanks to our current supporters!
FOUNDING ORGANIZATIONS
R·I·T
IS&T
PRIMARY SUPPORTING ORGANIZATIONS
NEH
INSTITUTE of Museum and Library SERVICES
The Andrew W. Mellon Foundation
FINANCIAL CONTRIBUTORS
CREATIVE memories
SAKURA

http://www.archivaladvisor.org/shtml/art_presdigmem.shtml

Proposed Service: **UnlockMe**

- Tools and instructions to overlay hosted services
- **Tell/show** people how to extract their content or..
- **Do it for them**, upon request, using
  - APIs
  - Screen scraping
  - Web crawlers
  - Scripts to do it the dumb way (machine equivalent of “Save” at file level)

Thank you!