On the Internet, Things Never Go Away Completely

Where is My Data and How Do I Really Know?

Thomas P. Keenan, FCIPS, I.S.P.
Faculty of Environmental Design
University of Calgary
2500 University Drive NW
Calgary, AB T2N 1N4
CANADA
keenan@ucalgary.ca

Abstract. The problem of persistent information residing on computer systems, and getting "into the wrong hands" has existed in theory since the first computer systems. In practice, it became a public issue when the emails of Oliver North and Bill Gates were introduced in court proceedings, and when Delta Airlines fired a flight attendant for her in-uniform blog posting. In a significant way, the digital trail that we leave behind is becoming a new form of "online identity," every bit as real as a passport, driver's license or pin number. The advent of new technologies, from Second Life avatars, to camera phones to video sharing sites, gives the question of "Where is My Data and How Do I Really Know?" some new and frightening dimensions. Future developments like "signature by DNA biometric" will make the issue more urgent and more complex. We will require new policies, technical tools, laws, and ethical standards.

1 Introduction

Even casual computer users know that simply deleting a file from their computer may not completely erase the data from the machine's disk system. While the file may become invisible to application programs, data clusters often remain, awaiting reallocation, and open to unauthorized inspection. Increasingly, additional copies of user data are found in slack space, swap files, recovery files, etc. Modern operating

systems are so complex that only a very sophisticated user would have any idea how to find and delete *all* copies of their data. Law enforcement investigators use this technical quirk to great advantage, pouring over seized computers with programs such as EnCase and FTK (Forensic Tool Kit.) The truly paranoid, or at least privacy sensitive users, often try to counter such sleuthing with programs such as Wipedisk, PGP Shredder and Evidence Eliminator.

The advent of the Internet has vastly complicated the whole problem of controlling user data. Search engine spiders, caching sites (both documented and hidden) mirror sites, web mail and web storage has led to a situation where, unless specific precautions are taken, one should essentially assume that data placed on the Internet can never be completely recaptured and may be viewed by others.

1.1 1.1 Historical Perspective – Single User and Timeshared Computers

In the earliest days of computer use, controlling user data was really not a problem. Scientists took turns using a computer on "booked time" and entered their programs either physically with wires and switches, or via removable media such as punched cards or paper tape. Output was either displayed on evanescent display screens or printed on a teletypewriter, so it could be torn off and taken away. When the author entered the world of computing, in 1965, instructions were clearly posted on the IBM 1620 computer console to zero out the entire 20,000 digits of memory before attempting to use it. This was good advice since the machine might "hang" or "loop" if it accidentally encountered improper data in memory. Reaching a block of zeros stopped the processor, allowing time for sober thought about programming errors. Anyway, we were so eager to run our own programs that it never occurred to us to snoop on the previous user's data.

The move to interconnecting computers raised the question of "where is my data?" to new levels. In the 1960s, the author worked on one of the earliest time-sharing systems (SHARER,) on a CDC 6600 computer at New York University. This system pioneered the concept of dividing up the power of a large (and then very expensive) mainframe computer among several users, and introduced the "exchange jump" instruction [1] which caused the computer to switch context between two users. A subsequent project carried out on a similar computer at the University of Calgary in 1972 demonstrated some of the vulnerabilities inherent in switching from one user to another. A prankster calling himself "The Missionary Unmasker" discovered other users' passwords and posted them around the campus. The author had to modify the operating system's code to clear out the relevant password fields between users.

1.2 1.2 Email as an Example of Vulnerability by Data Proliferation

Single-system email systems such as IBM's Professional Office System (PROFs) brought the issue of data deletion to the front pages of the world's newspapers. In the Iran Contra scandal, Reagan administration official Oliver North was embarrassed to find that PROFs emails that he thought he had deleted were produced as evidence. The matter went to several courts, and, according to a chronology [2] on White House emails, assembled by the Federal of American Scientists:

> "January 19, 1989…At 6:10 pm, on the eve of George Bush's inauguration, U.S. District Judge Barrington D. Parker issues a Temporary Restraining Order, prohibiting the destruction of the backup tapes to the PROFs system."

Other high profile instances of emails coming back to haunt the originator include the Jan. 5, 1996 memo from Microsoft chairman Bill Gates that was introduced as evidence in the company's antitrust trail. As reported by CNN [3] this email led to an interrogation of Gates about possible illegal business practices. And who could forget the posting, on the illmob.org website, of private phone numbers, photos, email addresses and notes belonging to celebrity Paris Hilton. (It is still unclear if this was done by social engineering or by a T-Mobile technical exploit such as the one posted at [4].) What makes that case particularly relevant is that, although illmob.org is a fairly obscure "hacker" website, the information rapidly proliferated to higher profile sites such as engadet.com and gizmodo.com.

IBM's ancient PROFs system had an interesting feature that many modern day email users would dearly love -- the ability to "recall" an ill-considered email message after it was sent. This was accomplished by simply deleting it from the delivery queue. Of course, if the recipient had already read, stored or forwarded the message, it was too late.

It's important to note that Jon Postel's original RFC 821 for SMTP (Simple Mail Transfer Protocol) [5] is silent on the issue of recalling mail, as is RFC 2821 which replaced it in 2001. [6] Some vestiges of this "unsend" concept remain in proprietary systems including Microsoft Outlook Exchange Server and AOL, but it's increasingly considered an archaic idea. It may well be impossible to implement now because of technical issues involving POP3 and IMAP servers, the use of web mail systems like Hotmail and Gmail, and a nasty security issue involving bogus recall requests that is described on www.whynot.com [7]

1.3 Web pages Have Become a Treasure Trove of Information

The introduction of the Mosaic web browser in 1993 caused a flood of Internet activity. Now, it would be unthinkable for a major company not to have a webpage. Yet those web pages may contain seeds of the company's own destruction. In a simple experiment, taking less than two minutes, high quality images of the corporate logos of the "big six" banks in Canada were obtained from:

- http://www.cibc.com/ca/img/default-logo.gif
- http://www.tdcanadatrust.com/images/TDCTLogo_big.gif
- http://www4.bmo.com/vgn/images/ebusiness/logo_financialgroup.gif
- http://scotiabank.com/static/en_topnav_logo.gif
- http://www.nbc.ca/bnc/files/bncimage/en/2/im_logo.gif
- http://www.rbcroyalbank.com/banners/oce/logo_rbc_banknng.gif

It should come as no surprise, then, that criminals preparing "phishing" schemes have little trouble creating very credible looking bogus bank web pages. In fact, they have reached the level of sophistication where the majority of their fake page is actually the real, functional code of the bank, with only a small portion of fraudulent content. It's also worth noting that, barring a significant change of their names and/or

logos, (which for marketing reasons almost never happens,) once these images are available they will remain usable practically forever.

The “Wayback Machine,” found at www.archive.org is an obvious example of unintended webpage archiving. Surely the system administrators of 1996 never intended their work to be easily viewable a full decade later!

2 The Present State of Data Persistence on the Internet

2.1 2.1 Data Storage by Government Agencies

This is an area shrouded in some mystery. Rumors describe vast disk farms in basements near Washington, D.C. archiving every email, web page change, Usenet post and even conversations by VoIP telephony. Internet users in China experience strange delays and “page not found” messages that lead them to believe they are being watched online. While the exact current state of such surveillance technology is not publicly available, much can be learned by reviewing the history.

First there was ECHELON, a secretive and controversial system operating by a number of governments to intercept and analyze communications of interest. It was publicly discussed in an article by Duncan Campbell [8] where he details various Signal Intelligence projects operating in the UK and the US, with code names like MOONPENNY, VORTEX and BIG BIRD.

Then came the US Federal Bureau of Investigation’s CARNIVORE system, which became public knowledge in 2000. According to an internal FBI memo, obtained, in censored form, under the Freedom of Information Act by the Electronic Privacy Information Center [9] “Carnivore was tested on a real world deployment (CENSORED)…This PC could reliably capture and archive all unfiltered traffic to the internal hard drive.”

The general consensus is that the FBI and its partners eventually replaced Carnivore with commercially available tools. This trend is consistent with the author’s own experience with another law enforcement agency. It is reasonable to assume that even better tools for data capture have been developed in the intervening years, and are now being deployed. It is also worth noting that the cost of data storage has plummeted, allowing the archiving of vast amounts of information at very low cost.

For many years, Usenet news groups were of special interest to governments and law enforcement because they were used for many questionable purposes, from trading pornographic images (legal and illegal) to planning drug deals and terrorist activities. That Usenet groups have been the subject of governmental attention is indisputable. According to a report prepared by the Electronic Privacy Information Center [10]

CompuServe, an on-line service of H&R; Block, based in Columbus, Ohio, removed from all of its computers more than 200 Usenet computer discussion groups and picture databases that had provoked criticism by a federal prosecutor in Munich.8 (The "banned" newsgroups were still available to CompuServe users who

used the service to connect to computers that carried the newsgroups. Information on how to do this circulated quickly through the CompuServe system.) Three days later, the Chinese government echoed the Germans' actions by calling for a crackdown on the Internet to rid the country of pornography and "detrimental information."

2.2 2.2 Data Storage by Companies and Individuals

Whether or not any governments were systematically monitoring Usenet group postings is somewhat moot, because they can just go do their data mining right now in a number of Usenet archives. The most famous was DejaNews, which allowed anyone to retrieve old postings. The author once accidentally embarrassed a teaching assistant by searching her name on DejaNews, only to find some fiery and radical political postings. They weren't actually her views, she pointed out; she was just trying to "infiltrate" a radical group to do an anthropology paper. Aside from the ethical questions there, the fact is that her (rather distinctive) surname remained attached to what may be an illegal (because of incitement to violence) posting.

DejaNews was bought by Google in 2001 and rolled into Google Groups. It contains postings back to 1981 (some with earlier dates like 1971 are undoubtedly the result of incorrect date setting) on predictable subjects like "Star Trek." One has to wonder if Chip Hitchcock, now a Fellow of the New England Science Fiction Association, would want to be reminded that in 25 years ago someone bearing his name wrote this:

Date: 17 Jun 1981 10:40:32-EDT
From: cjh at CCA-UNIX (Chip Hitchcock)
…Certainly her proportions were extreme enough to satisfy most people; was it that she refused to do a nude scene (which I find thoroughly unlikely for an unknown in present-day filmmaking)? …And do you think that one mark of a good actress is willingness to strip for the camera?

Yet it's up there, in Google Groups, for all to see. And it probably always will be.

3 Emerging Threats

There are many, many ways to let data out, and essentially (except for encryption or some kind of encryption-based "data expiry" and "rights management" schemes) no effective way to get it back. So it is prudent to consider the data proliferation risks inherent in new technologies, and how they may affect us.

Observers of young people born between 1980 and 2000, have commented that "for Generation Y, communication is all about MySpace and Facebook." [11] One might add that it's also about blog postings, sharing videos on YouTube, Instant Messenger Chat and phone-to-phone SMS messages. While the seemingly ephemeral nature of such communications might seem to minimize the risk of data dissemination and persistence, actually the opposite is true. Briefly, here are some of the emerging issues:

3.1 IM logging

Chats are now routinely logged on the computers of both parties. This provides an opportunity for unauthorized parties to read them, unobtrusively, at a later date. They can also be sent by email, and in fact, in Google's Gmail system, chat entries that occur while you are offline are automatically sent to you by email. So all the data persistence problems of email are becoming replicated in the chat universe.

3.2 Video sharing

Despite the intention of sites like YouTube to force viewers to watch videos in real-time, there are numerous free available programs to store them (KeepVid, YouTube Downloader, SnagIt) as well as the option of simply connecting the video stream via hardware to a device such as a DVD Recorder.

Every day, YouTube and similar sites receive numerous "takedown requests" from copyright holders and those who find particular videos offensive or invasive of their privacy. There is a formal procedure for handling these applications, as well as a process for getting a video re-posted if in fact it should not have been taken down under the company's policy. YouTube's broadly written "inappropriate content" clause [12] mentions material that is "unlawful, obscene, defamatory, libelous, threatening, pornographic, harassing, hateful, racially or ethnically offensive, or encourages conduct that would be considered a criminal offense, give rise to civil liability, violate any law, or is otherwise inappropriate."

Some videos just keep re-appearing and causing problems. According to Rabbi Abraham Cooper, Associate Dean of the Simon Wiesenthal Center, [13] a Nazi propaganda film called "Hitler Builds a Village for the Jews" is frequently re-posted on video sites by Holocaust deniers, forcing repeated takedown requests. The major video posting sites are now implementing "digital signature" technology to assist in automating the takedown process, but new video posting sites keep springing up all over the world. Some of them don't have the same level of scrutiny as Google-owned YouTube.

3.3 Blog Sites

Delta Airlines became famous, in a negative way, for firing flight attendant Ellen Simonetti "for posting inappropriate pictures (of herself) in uniform on the Web." [14] Many other bloggers have suffered in real life because of their virtual lives. Blogspot, created by Pyra Labs and acquired by Google in 2003, stores blog entries on Google's servers. According to Google's Privacy Policy for this service [15], "If you delete your weblog, we will remove all posts from public view." However, it goes on to say that "because of the way we maintain this service, residual copies of your profile information and other information associated with your account may remain on back-up media."

3.4 Skype and other VoIP products.

In its Privacy Policy [16] Skype distinguishes between your Personal Data (name, address, billing information;) Traffic Data (who you call;) and Communications Content (actually voice or data transmitted.) They of course note that they may be obliged to disclose any or all of these to law enforcement officials upon lawful request. However Skype also reserves the right to "share your Personal and Traffic Data with carriers, partner service providers and/or agents, for example the PSTN-VoIP gateway provider, distributor of Skype Software and/or VoIP Service and/or the third party banking organization or other providers of payment services."

Vonage [17] has a substantially similar privacy policy but also includes this warning about VoIP communications, "…no system or service can give a 100% guarantee of security, especially a service that relies upon the public Internet. Therefore, you acknowledge the risk that third parties may gain unauthorized access to your information when using our services."

3.5 Facebook, MySpace, Nexopia

Facebook suffered a major user backlash in 2006 when it launched new features called NewsFeed and MiniFeed. These programs sent all Facebook users information about the activities of their friends. An online protest group called "Students Against Facebook Newsfeed" was launched and attracted over 300,000 members, and the company modified its policy somewhat.

Most Facebook account holders believe that when they delete something (a wall posting, a photo, a compromising video) it's gone. But Facebook's own privacy policy (which few users have probably read) states "You understand and acknowledge that, even after removal, copies of User Content may remain viewable in cached and archived pages or if other Users have copied or stored your User Content." [18]

In any case, it is dead easy to right click on an interesting Facebook photo, capture a video, or make note of personal information provided when something is offered for sale in Facebook Marketplace. There's a good reason why certain law enforcement officials refer to it as StalkerBook.

MySpace, and Nexopia, provide free accounts to anyone who says they are 14 years of age or older. There is some human review to ensure that absence or truly offensive images are not posted. Some fairly intimate personal details are requested, and freely given, though perhaps not always with 100% honesty.

A recent Nexopia search displayed several hundred Calgarians who list themselves as being between 14 and 17 with "homosexual" as their sexual orientation. Most have photos and many have some personal information attached in blog entries. The site also lists the nicknames of their friends, allowing for social network profiling. Of course, many of these boys and girls are just amusing themselves, but they run the risk of information they disclose voluntarily on Nexopia causing them embarrassment and perhaps even serious problems later in life.

3.6 Second Life and other Virtual Worlds

Virtual worlds are nothing new, dating back at least to The Palace, that legendary virtual reality community created in 1996. It introduced many people to the idea of avatars, and conversing in a virtual world through chat bubbles. Now, Second Life claims to have 7.5M "residents" with 1.6M of them logging on in the last 60 days. There are virtual products and services, virtual real estate, and the ability to exchange Second Life's internal currency (Linden dollars,) for U.S. dollars.

Like, Facebook, the Second Life privacy policy cautions against expecting privacy with respect to information you disclose in the virtual world, i.e. "Please be aware that such information is public information and you should not expect privacy or confidentiality in these settings." They also note that they permanently retain the "registration file" of former customers even after they have ceased to use Second Life. They are silent on what happens to your other digital data, but it's a fair bet that your fuzzy little avatar and online transactions will be sitting on at least one backup file somewhere on the planet.

Ironically, the major concern about Second Life and similar systems may be the non-persistence of your data. As one writer recently noted in an online trade journal [19], "There are no standards that let you move your avatar, your virtual shop, or any of your innovations between virtual realities…if Linden goes down or bust, what happens to your Second Life shop?"

3.7 RFID and Bluetooth data

An experiment [20] at the MIT Media Lab demonstrated that Bluetooth-enabled cell phones produce enough data to track the movements of individuals as well as determine who they are spending their time with. RFID tags have been controversial, with the Brittan Elementary School in Sutter, California seeking to have all children tagged and parents opposing it on privacy grounds. [21] The use of the RFID in passports is also highly contested for reasons of privacy and security. [22]

3.8 Things We Haven't Invented Yet

A consideration of data retention should at least contemplate future technologies. As just one example, it is entirely conceivable that we will soon be signing documents and authorizing online transactions using biometric data, perhaps even our DNA signature. Very few jurisdictions have comprehensive laws governing the handling, storage, exchange and sale of biometric data. Aside from its highly personal nature and status as an identifier with non-repudiation characteristics, genetic data may also disclose health information about the subject and even other family members. This, in turn, could have adverse consequences in areas such as health care, employment and insurance.

4 Conclusion: Setting a Balance

Whether through government snooping, corporate data retention, personal hoarding or just plain accident, more and more of our data is being permanently stored away. Much of it can be traced back to us, either by name, IP address, or pseudonym. As storage cost goes to zero, there will be no technical or economic reason to ever delete anything. In fact the human cost of figuring out what to delete already exceeds the cost of buying another 500GB hard drive for most people. So we keep everything.

A related consideration, well beyond the scope of this paper, is that image and video search engines, and search technology in general, keep getting more effective. Not only will there be an embarrassing thirty year old video clip of you out there; anyone will be able to find it armed simply with a current photo of you and "reverse aging" software!

Governments and companies that deal with the public will need to continually re-consider their policies on data use and retention. All of us should think carefully about every word, video and photo that we put into cyberspace. "Would I want my mother or my next employer to see this?

If we don't set smart policies as a society, we might find ourselves moving in the rather Luddite direction suggested by a company called AlphaSmart. They're capitalizing on the fears of parents about their kids being online, and possibly leaving behind some digital footprints by selling the "Neo laptop." It's a computer with "versatile learning software for developing writing, keyboarding and quizzing skills." But, as their online brochure [23] explains, "Neo purposely does not include Internet capabilities. Students stay on task without Internet distractions — Web surfing, online games, or instant messaging."

It's not clear if the Neo is named as some sort of tribute to the Keanu Reeves character in the "Matrix" movies. Whether it is or not, it points to the fact that we all need to see beyond the illusion that our data goes away when we think it does. It's time to prepare intelligently for a world where everything we ever say, do, or perhaps even, think, may someday come back to haunt us.

5 References (All online citations accessed June 24, 2007)

1. Los Alamos Scientific Laboratory, Semiannual Atomic Energy Commission Computer Information Meeting, May 20-21, 1968, report LA-3930-MS, available online at http://www.fas.org/sgp/othergov/doe/lanl/lib-www/la-pubs/00320743.pdf
2. Federation of American Scientists, White House Email Chronology, http://www.fas.org/spp/starwars/offdocs/reagan/chron.txt
3. CNN, "Gates Deposition Makes Judge Laugh in Court," Nov. 17, 1998, available at http://www.cnn.com/TECH/computing/9811/17/judgelaugh.ms.idg/
4. Rootsecure.net, http://www.rootsecure.net/?p=reports/paris_hilton_phonebook_hacked
5. Postel, J.B., Simple Mail Transfer Protocol, http://www.ietf.org/rfc/rfc0821.txt
6. Klensin, J., ed., Simple Mail Transfer Protocol, http://tools.ietf.org/html/rfc2821

7. http://www.whynot.net/ideas/902
8. Campbell, D., They've Got It Taped, New Statesman & Society; Aug 12, 1988, pg. 10
9. Federal Bureau of Investigation, Memo, Case # 268-HQ-1092598, June 5, 200, UNCLASSFIED, posted on www.epic.org/privacy/carnivore/test_6_ 00.html
10. Electronic Privacy Information Center, Human Rights Watch, Vol. 8, No. 2, May , 1996, http://www.epic.org/free_speech/intl/hrw_report_5_96.html
11. Holland, A., Does Generation Y Consider Email Obsolete? http://www.marketingsherpa.com/article.php?ident=30010
12. http://www.youtube.com/t/terms
13. Cooper, A., Simon Wiesenthal Center, Private communication, May, 2007
14. Simonetti, E., "I Was Fired for Blogging," CNET News, Dec 16, 2004, http://news.com.com/I%20was%20fired%20for%20blogging/2010-1030_3-5490836.html
15. http://www.blogger.com/privacy
16. http://www.skype.com/intl/en/company/legal/privacy/privacy_general.html
17. http://www.vonage.com/help.php?lid=footer_privacy&article=399
18. http://ucalgary.facebook.com/policy.php
19. http://opinion.zdnet.co.uk/leader/0,1000002208,39287486,00.htm
20. http://reality.media.mit.edu/researchmethods.php
21. Electronic Privacy Information Center, "Children and RFID Systems," http://www.epic.org/privacy/rfid/children.html
22. Zetter, K., "Feds Rethinking RFID Passport," Wired, online edition, Apr. 26, 2005, http://www.wired.com/politics/security/news/2005/04/67333
23. http://www.alphasmart.com/k12/K12_Products/neo_K12.html