On the Internet, Things Never Go Away Completely

The growing problem of Internet data persistence

Thomas P. Keenan

Faculty of Environmental Design, University of Calgary
2500 University Drive NW Calgary, AB T2N 1N4
Canada
keenan@ucalgary.ca

Abstract. The problem of information "getting into the wrong hands" has existed since the first stored data computer systems. Numerous companies and government departments have been embarrassed by data left on un-erased media such as magnetic tape and discovered by inquiring minds. The advent of data communications brought the problem to a whole new level, since information could be transmitted over long distances to places unknown. The phenomenal rise of the Internet elevated the problem of Internet Data Persistence (IDP) to a public issue, as the "private" emails of public figures such 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. New technologies, from virtual worlds, to camera phones to video sharing sites, give the question of "Where Has My Data Gone and How Do I Really Know?" some new and frightening dimensions. Future developments like "signature by DNA biometric" will make the issue even more urgent and more complex. Coping with it will require new policies, technical tools, laws, and ethical standards. It has even been suggested that a whole new profession, sometimes called the "e-scrubber," will arise to assist in tracking down and deleting unwanted online remnants.

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 have 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 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 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, conducted by the author and taking less than two minutes, high quality images of the corporate logos of the "big six" banks in Canada were obtained from in June 2007 :

- 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_bankng.gif

A repetition of this experiment five months later, again using the Firefox browser, disclosed that, while some of the image locations had changed, they were all obtainable by simply right clicking on the appropriate bank's webpage and clicking

"Save Image As." 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.

Though banned by some laws, such as the UK's anti-fraud statute that came into force in 2007 [8] "Phishing Kits" remain for sale in numerous online venues. These give even an inexperienced, non-technical user the tools to create false websites and launch large volumes of spam. Various techniques such as Anti-DNS Pinning can be brought into play to make the recipient of these emails think that they are in fact coming from the legitimate website of a financial institution.

Another significant development is the advent of "Google Hacking," which uses the leading search engine to find information that was never intended to be found, including passwords, internal printer addresses, even logs of security vulnerabilities produced by commercial security scanning products. There are excellent online references such as [9] to explain Google Hacking.

Lest it be thought that all Internet Data Persistence (IDP) is connected to malicious computing, there are countless examples of innocent archiving, which may still have embarrassing consequences. The "Wayback Machine," found at www.archive.org is an obvious example of unintended (to the webpage creator) archiving. Surely the webmasters of 1997 never intended to be judged by their old work which is easily viewable a full decade later!

2 The present state of data persistence on the Internet

2.1 Data storage by government agencies

This is an area shrouded in some mystery. Rumors and urban legends describe vast disk farms in basements near Washington, D.C. archiving every email, web page change, Usenet postings 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. There is an excellent report on this subject from the Open Net Initiative which portrays the Chinese surveillance situation, at least as it existed in 2004-2005. [10] Based on actual testing, this report notes that "China's Internet filtering regime is the most sophisticated effort of its kind in the world. Compared to similar efforts in other states, China's filtering regime is pervasive, sophisticated, and effective. It comprises multiple levels of legal regulation and technical control. It involves numerous state agencies and thousands of public and private personnel. It censors content transmitted through multiple methods, including Web pages, Web logs, on-line discussion forums, university bulletin board systems, and e-mail messages."

Many other governments have done some form of clandestine monitoring of the Internet. One early example is ECHELON, a secretive and controversial system operated by a number of governments to intercept and analyze communications of interest. It was publicly discussed in an article by Duncan Campbell [11] 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 [12] "Carnivore was tested on a real world deployment [deletion] having recently come back from a deployment...This PC could reliably capture and archive all unfiltered traffic to the internal hard drive (HD) at [deleted]." 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 [13]:

"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." 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 their country of pornography and "detrimental information."

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 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 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." [14] 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. The line between telephones and computers is also being blurred. SMS messages can be sent from computers using sites such as www.blueskyfrog.com.au, which links to certain mobile telephony providers in Australia. Whether such data is being logged at the computer, the cell phone, or somewhere in between is an interesting question to which most people don't know the answer. Few people realize that their (anonymized) search queries are being displayed on giant screens in Google's California headquarters, as well as on websites like www.metaspy.com.

The November 2007 announcement by Google and partners of the Open Handset Alliance, based on open source technology, will further eliminate the distinction between computing and telephony. If you use your phone to access Google Maps there are ample places that might retain the details of just where you were going.

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 [15] 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, [16] 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." [17] 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 [18], "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."

That, in itself, is an understandable consequence of the technical architecture of the system. However, many aspects of that residual information are poorly defined. Who has access to it? Can it be subpoenaed? Can law enforcement just drop by and take a look? How long will it be retained? Cities that have installed surveillance

cameras in public areas have needed to wrestle with these problems. However, private companies have much greater leeway in crafting and enforcing their privacy policies.

Blogging has taken an interesting public twist, politicians and political candidates now using this technique to "get closer to the voters." Bill Clinton has a blog, and uses it to talk about his recent trip to Africa. Barack Obama's site, www.barackobama.com, features a "group blog" written by campaign staffers. The risk of course is that their words may come back to haunt them. Old election promises may be archived and resurrected. Speeches given to a group of students may be compared against those given to senior citizens. The net result may be more transparency. It may equally result in more obfuscation and even more oblique speeches.

3.4 Skype and other VoIP products.

In its Privacy Policy [19] 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 [20] 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 Social networking (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." [21]

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 ruefully refer to it as "StalkerBook."

The public's awareness of Facebook privacy issues has been raised by a provocative video "Do you have Facebook" posted on YouTube and now viewed over 125, 000 times. Essentially, it is a reading of the Facebook terms of service, combined with a conspiracy theory about possible links between the site and certain US government agencies. As noted in the video, "all of the above raises more questions than answers."

MySpace, and Nexopia, provide free accounts to anyone who says they are 14 years of age or older. There is some review by human moderators to ensure that obscene or highly 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 Unsolicited data collection (ChoicePoint and ZoomInfo)

ChoicePoint (Alpharetta, GA) is one of the largest data brokers in the world. It collects personal data ranging from social security numbers to real estate holdings, and is not above sending people into courthouse basements to copy out divorce judgments. According to an online trade press article "it also offers businesses, government agencies and nonprofit organizations software technology and information designed to anticipate and respond to economic and physical risk, and it analyzes information for the insurance sector. Its database contains about 19 billion records." [22]

Most people in that database didn't ask to be there, and may well be unaware that they are. Whole government agencies such as the New York City's Office of Vital Records, have outsourced birth and death record processing to VitalCheck, a ChoicePoint company. Whether or not they guard the personal information with the same care as a government office is, of course, open to debate. It is known that when the Alberta provincial government moved Drivers License processing to private sector vendors, serious security flaws such as fake driver's license on official license blanks were reported.

An even more subtle form of unsolicited data collection is typified by the site www.zoominfo.com. This company has data on almost 39 million people, much of it obtained by scanning the Internet for web pages, press releases etc. The vast majority of the records are described as "automatically generated using references found on the Internet" and "This information has not been verified." When I checked my own profile, which had a great deal of correct information, I was surprised to learn that I was on the board of a defense contractor that I had never heard of.

3.7 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 [23], "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?"

3.8 RFID and Bluetooth data

An experiment [24] 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. Researchers outfitted willing subjects with "always on" phones that could discover each other and log precise locations through GPS technology. The findings included the concept of "familiar strangers," people you are often near but do not actually know, and suggested that it might benefit a company to introduce them to each other.

While this data trail was purely voluntary, one could easily imagine devices of this nature being used for social control purposes such as keeping track of who met with whom in a hallway at a busy conference. Indeed at the 2003 World Summit on the Information Society (WSIS) that took place in Geneva, Switzerland, rumors were rampant that delegates were indeed being tracked electronically through the RFID tags embedded in their badges. There have been numerous accusations that the procedures for handling personal data at WSIS 2003 may have violated legislation including the European Union Data Protection Directive. [25]

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. [26] The use of the RFID in passports is also highly contested for reasons of privacy and security. [27]

3.9 IP and MAC address logging

An underappreciated aspect of digital trail we leave is the logging of addresses such as our Internet Protocol (IP) and MAC (Media Access Control) addresses. The former is assigned by an Internet service provider and can vary over time. However, as more people move to services that assign IP addresses for a long time (such as cable providers) the IP address becomes more useful for tracking. As one example, if you forget your password on the Second Life site, your IP address is included in the email reminding you of your password. This is intended to provide tracking information for bogus password reset requests, but it also provides some degree of tracking information on innocent people. In fact, in a court case in Alberta, Canada, evidence was introduced in an "Internet defamation" case that included the IP addresses used to post on stock discussion sites.

MAC addresses are unique to individual Network Interface Cards, so they would seem to be the perfect identifier for a machine, and to be a dream come true for those trying to do computer forensics and trace people on the Internet. However, because of the design of TCP/IP, the protocol underlying the Internet, MAC addresses are only transmitted up to the Data Link Layer so they are not generally available across routers. The net effect is that retrieving the MAC address of a remote computer is generally only possible if it is on the same host. Also, there are techniques for "spoofing" MAC addresses. MAC addresses are sometimes used for the generation of license keys for proprietary software that is authorized to run only on a specific computer. A mathematical comparison is made of the computer's MAC address to what it is supposed to be, as encoded in the license key.

3.10 Public and shared computers

Who hasn't used an Internet café, or a hotel business center's computers or those in an airport lounge? We typically do that without regard for the fact that we have no control over the hardware, software or network that we are using. What better target for unscrupulous hackers than Business and First Class passengers?

There are numerous exploits that could be placed on public computers. A press article [28] reported that malicious software ranging from keystroke loggers to Back Orifice was reported on airport lounge computers. Even without hacker attacks, there are some simple problems relating to things like sending email using a common product. "Outlook Express is probably not configured to allow emails to be sent from these machines, so any message created simply moves to the system's 'outbox' where it remains indefinitely after the user clicks 'send'." This allows the next user to come in and review those messages. In fact, unless you explicitly clear your "Sent Mail" folder before leaving the airport computer, you are probably leaving all that juicy information in there as well.

3.11 Things we haven't invented yet

A consideration of Internet Data Persistence should 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.

Other trends include the phenomenal decline in storage costs, leading many to believe that the costs will approach zero and all data will be retained because it is uneconomical to get rid of it. Couple this with ever-improving search technology, pioneered by firms like Google, and the ability to find the proverbial "needle in a haystack" and to embarrass a company with it, is a very real threat.

3.12 An emerging profession: the e-scrubber

In a provocative blog posting [29] from the firm Social Technologies, a number of "New Jobs for 2020" were postulated. One was the E-scrubber, who "Works to undo or minimize the indiscretions that people accumulate on the Web." Given the volume of embarrassing photos, off color jokes and other content posted every day, it seems that this might indeed be a growth industry. Of course, the same tools that would allow e-Scrubbers to track down indiscretions could, in the wrong hands, be used to find them.

An even bigger question arises, "How Would the E-scrubber Know When the Job is Done?" After all, there are many "Deep Web" Internet databases that are not visible to the general public, and not indexed by search engine spiders. There could always be one more place where a given piece of data is stored, perhaps in an encrypted form. So, the real answer to the question is not "the Job is Finished" but, sadly, "We've Done All We Can for You."

4 Conclusion: setting a balance before the technophobes do it for us

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.

The problem is exacerbated by the relentless improvement of search engine technology. Soon, 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!

The risks are very real, and no-one is immune. Even the United States Air Force was a victim of Data Persistence, leaving sensitive data on un-erased magnetic tapes that were sold as surplus. [30]

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, asking "would I want my mother or my prospective employer to see this?

If we don't set smart policies as a society, we might find ourselves moving in the technophobic 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 [31] 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 heroic hacker Keanu Reeves plays 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.

References[1]

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. http://www.opsi.gov.uk/acts/acts2006/pdf/ukpga_20060035_en.pdf
9. http://johnny.ihackstuff.com/
10. Open Net Initiative, "Internet Filtering in China in 2004-2005: A Country Study, http://www.opennetinitiative.net/studies/china/, accessed Dec. 7, 2007.
11. Campbell, D., They've Got It Taped, New Statesman & Society; Aug 12, 1988, pg. 10
12. Electronic Privacy Information Center, press release, November 16, 2000, http://www.epic.org/privacy/carnivore/11_16_release
13. Electronic Privacy Information Center, "Silencing the Net – The Threat to Freedom of Expression On-line, Human Rights Watch, Vol. 8, No. 2, May, 1996, http://www.epic.org/free_speech/intl/hrw_report_5_96.html

[1] All online citations accessed June 24, 2007 except as noted

14. Holland, A., Does Generation Y Consider Email Obsolete? http://www.marketingsherpa.com/article.php?ident=30010
15. http://www.youtube.com/t/terms
16. Cooper, A., Simon Wiesenthal Center, Private communication, May, 2007
17. 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
18. http://www.blogger.com/privacy
19. http://www.skype.com/intl/en/company/legal/privacy/privacy_general.html
20. http://www.vonage.com/help.php?lid=footer_privacy&article=399
21. http://ucalgary.facebook.com/policy.php
22. Campanelli, M, "Checkpoint to Divest Three Units," DMNews, July 13, 2006, http://www.dmnews.com/cms/dm-news/database-marketing/37474.html, accessed December 9, 2007.
23. ZDNet, "Virtual Worlds, Real Problems," June 11, 2007, available online at http://news.zdnet.co.uk/leader/0,1000002982,39287486,00.htm, accessed Dec. 9, 2007.
24. http://reality.media.mit.edu/researchmethods.php
25. http://europa.eu.int/comm/internal_market/privacy/index_en.htm
26. Electronic Privacy Information Center, "Children and RFID Systems," http://www.epic.org/privacy/rfid/children.html
27. Zetter, K., "Feds Rethinking RFID Passport," Wired, online edition, Apr. 26, 2005, http://www.wired.com/politics/security/news/2005/04/67333
28. http://www.theregister.co.uk/2005/09/21/airport_pc_security_lax/
29. http://wcpl-businessbriefs.blogspot.com/2007/09/new-jobs-for-2020.html, accessed October 27, 2007
30. Neumann, P., "Illustrative Risks to the public in the use of computer systems and related technology," ACM SIGSOFT Engineering News, Vol. 21, No. 6, pp. 16-30, 1996.
31. http://www.alphasmart.com/k12/K12_Products/neo_K12.html