Knowledge Everywhere: The Distributed Memory of Social Media

Alexander Halavais
Quinnipiac University

Archived surveillance

Anyone who has watched police procedurals knows that most store surveillance cameras, when they are working at all, recycle the video tapes. So, you might have a recording of an incident right after it happens, but going back to see what occurred two weeks or two months ago is more difficult. The H.264 compression algorithm has become the standard for new digital surveillance systems[1], and hard disc sizes in excess of a terabyte are common, providing a video history of several months. Often, these systems provide for further back-up via USB ports or over the internet. Even as storage is increased and compression technologies are made more efficient, the cost of such systems is reduced. At the most extreme end, component cameras like those found in mobile devices are becoming so inexpensive that they can be easily added to almost any electronic device. As a result, we have the infrastructure necessary to know more about our world today, and to be better able to recall our world of the past.

Our lives are being recorded on video to a greater extent than they have been in the past, but this more visible form of recording has already been predicted by the casual archiving of dataveillance. Electronic recordings—whether of video or of text—are slippery; they can be copied and moved as easily as they can be deleted or lost. It is worth asking whether the slippery nature of electronic texts, of the cloud, and of participatory media, also means that we should be looking at archiving in new ways. More precisely, should we be thinking of the internet as a kind of holographic memory, and exploring ways of systematically building archiving into it, and making those archives useful to future historians and present investigators?

Surveillance doesn't have to be a bad word

We are accustomed to think of surveillance as asymmetrical: the state or corporation watching the individual. Such a view is well-formed, as many of the technologies we associate with surveillance enhance a power differential. This asymmetry is built into the word "surveillance": to watch over. But this is not the only way in which technologies of observation may be used. Harold Lasswell, writing in 1948, argued that surveillance is one of the chief social functions of communication systems in society[2]. Those who guide social policy and steer the government must be well informed for society to work effectively. The CCTV camera monitored by the police seems to be a different kind of surveillance than that of the evening news camera capturing images of a political protest, but both cases represent opportunities to shape society, based on observing and sharing evidence.

More recently, individuals are capturing their own images and observations to be shared and used to shape perceptions and inform their peers. Historically, these may have been private images of interest mainly to friends or family. But the ubiquity of digital recording technologies, and the ability to easily distribute and share the material, has led to a new kind of surveillance, a surveillance on the edge of the public-private divide. Of course, traditional watchers—the state and the news media—have attempted to make use of these new forms of surveillance, urging people on the New York subways to "See something, say something," and requesting that those with interesting video send it to CNN's iReport website[3]. But many of the observations made by individuals are shared directly with wider publics on social networking sites, blogging platforms, or content hosting sites for video, audio, or images.

A number of neologisms for this sort of peer-to-peer surveillance have been proffered. Steve Mann has suggested "sousveillance" and others "equiveillance" or "panveillance"[4]. Many indicate Brin's phrase in this context, "reciprocal transparency," and his argument that the only effective response to an increasingly surveilled world is policy that increases the transparency of the watchers[5]. It may, however, be more useful to reform the idea of surveillance and recognize that it changes as society changes. Lasswell's discussion of the surveillance function is frequently associated with the role of the news media. In fact, there is some indication that Lasswell himself saw the role as much more determined by the structure of communication in society, and that while some form of "watching over" was necessary to its function, who is watching and how that is done may change over time[6]. Recent abilities to easily share personally-created media, and social structures that encourage such exchange, have combined to provide for a new surveillance function.

These new structures of media affect the way in which we will make sense of this age. In preliterate societies, the knowledge that survives is often recorded versions of events that were passed from speaker to listener until they reached someone able to record the story. Over the last millennia, we have been forced to rely on public, official accounts made initially by the government, and eventually by a growing news media, along with private correspondence and diaries. The emergence of a large-scale participatory media has created an influential channel through which we learn about what is important to our friends, and to a much larger social network. It effectively bridges the existing, private, and personal networks of surveillance with mass mediated forms of surveillance.

Will these relatively ephemeral and distributed communications, these conversations through personal media, be something that can be effectively shaped into a narrative? As one observer has recently noted "If web 2.0 was all about democratizing publishing, then the next stage of the web may well be based on democratizing data mining of all that content that is being published"[7]. With such a process of examining present mass, distributed conversation will extend backward as well, so that we may learn more about what people knew, and what they did with that knowledge. Today, marketers are already eager to quickly come to terms with these rapid, large-scale, distributed conversations. The question is whether such broad phenomena can be recorded for later retrospective analysis.

Archiving 2.0

Electronic media, and the web especially, pose a particularly difficult problem for archivists and historians. This includes everyone from librarians at national libraries to those responsible for document retention in large organizations. Digitization makes it easier to create and revise documents, which has led to an explosion in the amount of text and other forms of information available to record and preserve. But the preservation of this data, and the meta-data needed to make sense of it, is made much more difficult when it is not clearly embodied in a physical medium. And on the individualized web, production of media by large portions of the group formerly known as the audience has created even further problems.

Much has been made of the relatively ephemeral nature of modern electronic storage. The practice of storing CDs with relevant data has turned out to be less than ideal[8]. Recently, libraries that have received the collected "papers" of modern authors have faced the prospect of decoding data stored on varying forms of magnetic and optical tapes and disks. In many cases, recovering the data from these physical stores requires obtaining the drives it was originally created with, and interfacing these drives with modern systems. And things continue to grow even more ephemeral.

Unlike the writers of the previous decade, it is entirely possible that a modern author may not have any of their manuscripts stored on removable media. It may reside on a voluminous hard drive in their computer; or increasingly, distributed across a number of hard drives in server computers around the world. The use of cloud computing in document creation is becoming more common, and follows other web-based applications, including email. Some who keep all of their work on a local computer subscribe to internet-based backup services. The future document may be like the one you are reading: born on the internet, and never transcribed to paper or stored on other portable media. In this case, what needs to be preserved is not some particular physical embodiment of the text, but rather the text itself and some indication of its context—that is, the way it was presented to the reader when it was first distributed[9].

A more participatory media culture makes preservation more difficult. There may have been a time when there was a clearer line between those personages that expected to be remembered by historians and those who did not. The former—who were largely self-selected—had some idea that their words might be read after their own deaths, and took some measures to avoid discarding correspondence and drafts of their writings. Today, we are all authors in some sense, and perhaps some are developing a sense of leaving behind a trail of our existence. Warhol had us famous for fifteen minutes, but in the network society, we will all be famous among fifteen people[10]. Whether or not future historians will seek records of the average citizen, or will be more interested in the most influential among us, there is some feeling that we all carry the potential to become famous someday.

Unfortunately, as many have found, the web represents an extraordinarily challenging target for archiving. Even the simplest form of web publication, the equivalent of a brochure containing static images, can be a difficult target to copy and preserve. But particularly during an era of more dynamic, database driven sites, the collection and preservation process becomes even more difficult. New content is produced at an extremely high rate, and much of it disappears quickly as well. This constant churn, and the bandwidth required to collect such large amounts of material, means that it is impossible to get an accurate, synchronous "snapshot" of the content of any large subsection of the web. The volume of material alone is staggering, but added to this is a large and growing range of formats, including video and audio content.

Part of this is driven by new leakage from the private to more public—or at least transparent—realms. Content carried in blogs, microblogging and awareness platforms like Twitter, and social networking sites like Facebook was formerly unavailable to the public observer, often hidden behind the veil of email, telephone, and face-to-face communication. As some of this becomes more publicly available, it represents a rich source of data not only for archivists, but for marketers. It is an embarrassment of riches, and one that is too easily lost.

Even more private material is now archived privately as more applications store information on the server side. The data Google alone collects on its users through a myriad of web applications is staggering. Email, documents, photos, videos, and full backups of personal computers are often stored in the computing "cloud." Although these are not immediately available to the archivist, what happens to these materials later in the user's life, or after the user's death, remains an open question. When compared to the vast, unstructured data of the web, the even more vast structured, private collections of service providers represents an even greater challenge to archivists, especially given the legal and privacy concerns surrounding these collections.

Caught between these public and private sources of information are the semi-public spaces like Facebook, which, through its terms of service, provides at least some minimal barrier to public scrutiny. This group is particularly difficult because it remains unclear how private it is, how private its users expect it to be, and what sort of social controls it places on "friends" within the social network itself. The near ubiquitous use of Facebook and related technologies by the generation currently in high school suggests a future in which we, at least potentially, will be able to see a presidential candidate's favorite activities, political attitudes, and social connections from an early point in their lives. We mythologize these connections for many public figures, and perhaps none so much as our presidents. We might smile over Lincoln's awkwardness in early courtship, as revealed by his letters, but because he became president, we at least have these minimal documents from his early history. Imagine the same kind of record of communications preserved for every individual.

The initial focus must be on the public web, which holds its own challenges. The Internet Archive remains the first and best effort at archiving the web as a whole. But the Archive is hardly the only archive, or the only archive that seeks to include all public material. Google has created an archive as part of their aim to index the world's information. Such an index requires obtaining a copy of the material for analysis. As a result, in some instances the Google cache contains materials that might be missed by the Internet Archive, a result of a substantial investment in crawling infrastructure. But because Google's collection is incidental to its stated primary mission, public access to those materials is quite limited.

Google is not the only large service provider to have cached large portions of the web. At one point America Online cached portions of the web to increase the speed of access to web materials for its subscribers, for example[11]. Many private companies collect smaller slices of the web as part of their normal operations. Other pieces of information, including traffic data, are collected, sifted, and compared. Large hosts of content, from eBay to YouTube, no doubt have some archive of material that has been submitted in the past. These data may come to be of interest to archivists and future historians, though they are fraught with legal and privacy entanglements. Recent attempts by national governments to obtain these records demonstrate their value and the controversy surrounding them.

More recently, efforts have been made to collect subsets of the web, often among national archives and libraries. Charged with preserving as much of the national web as possible, librarians quickly encounter the question of selection. Although the earliest libraries, and especially the Alexandrian library, made an effort to collect all recorded knowledge, modern libraries have always had to weigh space and budget requirements against the desire for a complete collection, and have had to seek out the most important books and papers to preserve and make available. The question of importance of material on the web is in some ways similar to earlier problems, and in other ways wholly new. Libraries have been able to rely on other institutions, from universities to publishers, to provide an initial level of filtering, but the democratization of production on the web means that selection of material is more difficult. In many ways librarians have been faced with the same problem that faces search engines: finding the valuable and credible information within an overwhelming sea of content. Both must find algorithms useful for judging salience, though the search engine seeks to meet the needs of more immediate users, while archivists must provide materials that are as useful to researchers next month as they are to scholars of the coming millennia.

A second approach is to collect materials surrounding a particular event. The September 11th attacks on the World Trade Center and the shootings at Virginia Tech were quickly followed by an outpouring of different kinds of communication, much of which might have been easily lost if not preserved immediately. Likewise, national elections in the United States and elsewhere have become the subject of web archives. Such an approach has the advantage of providing at least the basis of a selection criterion, and thereby limiting the scope of the collection to something more manageable. The disadvantage, of course, is that it limits our collection to what are already considered to be events of social import, and may miss the concerns of everyday life online.

Each of these approaches, however, hints at some of the challenges faced by the web archivist. Perhaps the most difficult archiving task is obtaining and tracing elements of the social web, because it is a moving target. There have been efforts to capture and analyze large portions of the blogosphere, and these probably represent the best model for the rest of the web. The British Library framed the archiving project as “saving Shakespeare’s blog,” a phrase that suggests that the future value of a person’s blog may not be knowable today. But a larger argument suggests that, in the aggregate, most blogs are worth preserving, because they reflect the interests and productive work of a very large number of people from very different backgrounds. The issue of time in the collection of blog text is made even more acute in the case of “micro-blogging” sites like Twitter.

One of the issues that capturing blogs has raised is the degree to which we should attempt to capture context and structure. Blogs, and the social web in general, are highly intertextual, drawing together a dispersed conversation. An exploration of a single blog post should, to whatever degree possible, be contextualized within the larger blog and the contemporaneous web. It may also be valuable, particularly for automated forms of research, to be able to extract data related to the structure of a post: Who wrote it? When was it published? What tags or topics did the author use to categorize it? These are questions we can ask now about blogs, but the same questions can often be asked about the wider social web, and provide cues for future researches working through the archives.

While the above provides some indications of success, it is also frustratingly overwhelming. How could we ever capture the vast web to any extent that could be called representative? I propose that we give up on making archives, and instead focus our efforts on building tools that allow the web to be self-archived.

The one-one map

Jorge Luis Borges and Lewis Carroll both describe maps that are 1:1 scale, and therefore cumbersome and unsustainable[12]. The project of archiving the web feels analogous to such a map. In fact, because we hope to collect and derive “extra” metadata from the structure and context of these materials, it is perhaps even more extensive. One might suggest that you would need all of the computers and storage in the world to be doubled in order to store just a snapshot of the current contents of the web. This is not the case, since there is already significant redundancy built into the web, but the amount of storage and processing necessary would be tremendous, and would grow at the same rate that the web itself is now growing. If archiving the web seems like a Sisyphean battle, it is because there is no way such resources could be gathered to run a centralized copy of the web.

We might, for the moment, draw an analogy to some of the ways in which computers store information, and avoid losing important data. One way in which they do this is to run two disk drives in parallel, with one disk mirroring the other (RAID 1). When one of these disks fails, it is easily replaced with a new mirror, and no loss in data occurs. In some sense, Google’s cache of the web, though incidentally, probably represents the closest we will come to a mirrored copy of the web's content, but it is only synchronous, not archival, and it is far from complete. Moreover, the Google effort is extraordinarily resource intensive.

An alternative is building an error-correcting web, increasing the redundancy somewhat in order to create a distributed collection. Error correcting codes add a little piece of information to each record to make sure that everything "adds up" and are able to correct distortions or missing data. With a bit of distributed redundancy, we can avoid some of the resource and bandwidth hurdles that Google's approach requires. It also seems appropriate to use distributed, small-scale approaches to archiving a structure like the web, which is already distributed and works through small-scale agreements.

You're soaking in it!

This sort of self-archiving is already happening and we should recognize and extend these existing practices where appropriate. At the most basic level, most of us already archive our materials in ways that we might not have in the past. The growing availability of disk space means that throwing out digital photos, emails, or document drafts is often more trouble than dropping them in an unused directory. Likewise, blogging systems and other forms of online distribution often store archives by default. So at some level, our traces are already better kept than they might otherwise have been, particularly for those of us who don't expect to become the subject of biographies in the future[13].

But beyond personal and organizational archiving, the machinery of the web is developing its own distributed memory. As already noted, Google caches web pages in a process that is incidental to their process of indexing, but there are other purposive caches as well. Flash crowds (often referred to as the "Slashdot Effect" or the "Digg Effect") overpower the servers of many web sites that become suddenly popular, making them suddenly unavailable to the potential audience. One way to avoid making these pages unavailable is to cache them on a server that is better able to handle the traffic. Since 2004, the Coral Content Distribution Network has created distributed caches of files that are in high demand, often due to flash crowds[14].

The Coral cache, along with those offered by Google and the Internet Archive, are in particularly high demand among users of the Digg social bookmarking site. On Digg, users discover content that has been endorsed by their fellow users, and like many such systems, the best endorsed sites are more likely to get even more endorsements. In a short period, a site may begin receiving thousands of visitors a minute. While large sites may be able to plan for this, Digg often encourages these large crowds to visit otherwise obscure sites, which may be on small servers with the capacity to handle only modest traffic. So it is natural that these sites come to have the most need for being cached, but quite by accident, these caches contain the material that a large subset of users found to be worth visiting, noting, and in some cases commenting on. It is a library built on the popularity of the materials found there, rather than on any other specific selection criteria.

There are other bookmarking systems—StubleUpon, del.icio.us, and Diigo, among them—that aggregate collective interest. Furl (recently acquired by Diigo) went one step further, combining the archiving process with bookmarking to provide users with a personal archive of pages of interest. Another source of a tacit selection function is the links from blogs: more popular content will receive more links from the blogosphere. This is equally true of microblogging platforms like Twitter, though in many cases the links from these sites are passed through URL shortening services. At present such services actually may make archiving more difficult, by obscuring the original site if the shortening service fails, but there are current efforts to crawl and archive the link relationships recorded[15]. The next step—archiving the actual page—is a relatively simple extension to this process.

Finally, there have been software tools available to the average web browser from the earliest days of the web that allowed them to save and store materials for later use. These included tools like "Web Historian" that installed as toolbars in browsers, and tools that allowed for copying and offline browsing. There are other approaches, crawling tools like "wget," for example, that can be used to mirror web sites, or Adobe Acrobat's ability to save the contents of a site in a non-interactive form. Tools built more recently on top of Google Gears provide for offline archiving of online site content, including RSS feeds aggregated in Google Reader. Finally, most browsers create a local cache of content to speed up return visits to the same website. What binds all of these local, client-side caches is that they are generally inaccessible to users on other machines. What would it mean if some of these pieces were linked together?

@rchive@home

There are two approaches of peer-to-peer archiving that might be fruitful: browser sharing and blog-based archiving. My effort here is not so much to argue that these should occur—though I do think they should and I hope to encourage them—but rather that they are so clearly venues for archiving that it would be surprising if they did not, sooner or later, lead to a distributed archive of the web.

The idea of peer-to-peer sharing of browser caches is not new. Iyer, Rowstron, and Druschel developed the Squirrel project at Microsoft Research to do just this at the beginning of the decade, and a number of others have followed suit[16]. Such systems are generally intended to counter flash crowds, as the more centralized caches have. For that reason, much of the research in the interim has addressed approaches to reducing latency, efficiently discovering neighbors, and the like. The idea that content may be saved for longer periods of time and in various versions has largely been left aside.

The disadvantage of such archives is that they may be deleted easily or made unavailable when the client computer is not available to the network. Another alternative is caching and self-archiving of links made from blog posts. Bloggers are already contributing to the distributed archive by archiving their own work. Many of these blogs act as filters, providing links to interesting news and information on the web. As blog archives get older, these links are particularly susceptible to "link rot." Since blog authors are interested in maintaining the integrity of their own blog's archives, it would not be difficult to create a system of archiving that builds on the existing blog, allowing for links to be cached locally. Since blogs are already on web servers, they can make these archived pages available to the public more easily, or may be exchanged across blog archives to provide for redundancy[17].

In some sense, this just moves the search problem toward the future. While these pages may be preserved, it hardly matters if they cannot be found. In the case of blog-driven archiving, such a system could easily produce an index of the pages on the server, which could then be harvested by specialized search engines. The result would be a collection of snapshots of the same site by different blogs around the web. As a growing number of sites move to the blogging format, this would represent a substantial, open, collection. Moreover, the most popular sites would likely be archived by the largest number of blogs, providing more protection to the sites judged most interesting by the collective wisdom of the blogosphere.

Set in stone

The traditional solution to the problem of electronic text has been to print it out and store it. Any magnetic storage medium is seen as dangerously unstable—the idea that such archives might be distributed and the location of physical archives indeterminate is, no doubt, troubling to archivists and historians. What little we know about collapsed civilizations comes to us via records that were literally set in (or on) stone. Moreover, the physical embodiment of a message has frequently been one of the few ways of determining its historical authenticity.

Of course, printing a large portion of the web is an untenable project. And for all of the advantages of a physical record, it is often an all-or-nothing affair. One of the reasons we have any records from antiquity at all is that they were, in fact, distributed archives. Only parts of those archives have survived, and our own assessment of archiving approaches is likely colored by a survivor bias. The fact that the Dead Sea Scrolls lasted as long as they did has a great deal to do with the way in which they were preserved and stored. But there were likely many more examples of well-intentioned but ultimately flawed archives that we simply will never know about.

The particulars of the Internet Archive—what sort of computers they are stored on and where those computers are physically located—mattered a great deal more before their archives were mirrored to their Alexandrian counterpart. With two copies of the archive, in different parts of the world, we feel more confident that they are less susceptible to either a cataclysmic physical event like an earthquake or a flood, or the more gradual demise of an operating system or computing platform. The archives, in digital form, are no longer as physically grounded. A fully distributed archive would be even more divorced from the hardware on which it is stored. The disadvantages are clear enough. When people decide to give up on blogging, for example, or forget to pay for hosting, some small part of the archive disappears. The existence of the archive may, at least to some extent, rely on the indefinite popularity of the technologies of the world wide web and perhaps the social currency of blogging, both of which are, most likely, temporary. Yes, the process of collecting material when it is blogged, and making indexes of that material publicly available, encourages mirroring and a second stage of harvesting.

But at some level, there needs to be an estimation that the substrate on which this archive is inscribed, the world wide web itself, will not disappear. If individual sites go away, there may be some diminution in the quality of the archive, but the integrity of the whole would remain. With a continuing and active group of users, the data would likely continue to be made available as new formats emerge. In other words, this is an archiving solution that works only as long as there is a world wide web. The advantage to this compromise is, with just a little initial work, a global network that very nearly archives itself.

Ownership and Privacy

This sort of distributed archiving faces many of the same kinds of copyright issues that face other forms of archiving. It is made more difficult, in this case, for several reasons. First, libraries fulfill a particular social role, a role that is sometime recognized in intellectual property law as different from the individual's. Under United States law, Internet Service Providers also can exercise special privileges. Even with these special protections, centralized caches of web pages—both commercial, like Google, and non-commercial, like the Internet Archive—have faced continuous challenges and questions with regard to copyright. An individual archiving system would likely continue to adhere to the robot exclusion protocol, but this alone is not enough to protect an individual archivist from the copyright holder.

The non-organizational status of the individual also places them in a relatively weaker position when it comes to fighting a copyright challenge, while larger organizations may be able to draw on greater resources. In some sense, this may actually be an advantage, in the aggregate. Attempts by the RIAA to crack down on file sharing by suing individual downloaders have been expensive and largely ineffective. However, the uncertainty of the legal situation may dissuade wide adoption of a distributed archiving effort.

An affiliated issue is one of authenticity and the ability to find the original document and for the author to be recognized for its creation. In particular, as with the use of URL shorteners, cached or archived copies make it slightly less likely that the original will be linked directly. This can also reduce what is commonly called the "Google juice" for a particular site, removing a source of traffic, as well as a tacit endorsement of the site that is used to determine its relevance. There are, at least potentially, ways of mitigating this problem technologically. Archives could be made less available, for example, until the original document has changed or been removed. But especially for obscure sites that might only be archived by a single blog, the process of archiving leads to some confusion.

Finally, there is the larger question of semi-public venues. The initial focus of the archive would be on materials available on the public web that would be preserved publicly. Even under these conditions, there are some areas in which the law and expectations are murky. Can users who publish on public—but unpopular—web pages be surprised when their work is preserved and re-presented in new contexts, and if that happens, what are the ethical obligations of the archivists? This is already difficult in the case of documents that are clearly public: ones that do not assert restrictive terms of service, and do not exclude web crawlers.

When it comes to comparatively less public sites, the question becomes even more difficult. Can we say that someone who has presented a profile on Facebook to a thousand of their "friends" really has a "private" profile? The question is moot, as Facebook makes clear that (for the moment) the materials on their site cannot be crawled and redisplayed elsewhere. Of course, our process of selection makes it unlikely that a link to Facebook would appear, since such a link would only work for a relatively small number of those who tried it, but it does raise the issue of how much is lost when we restrict ourselves to public materials, and how to address cases that are not clearly public or clearly private.

A medium to remember

Distributed archiving of the web does not directly address the question of search; search engines will be even more important as we collect a layered, holographic archive of the public web. But it does bring us back to the question of selection. An engine that provides a perfect copy of the web is impossible. Collections based on individual browsing habits (that is, the sharing of local browser caches) are likely to favor the most popular sites over the more obscure. The use of blog links—which often provide access to less popular material[18]— and social media as a selection tool for archiving still enforces a somewhat distorted view of the web and of the world. These collections will remain relatively vast and unstructured, requiring the use of search engines to effectively uncover archived material. This evades the question of selection by opting for non-selection, and pushing the problem to later researchers.

Or, at least it does in part. In the aggregate, blogs represent at least a subset of the *vox populi*, and the idea of leaving the selection of materials up to non-historians is galling for some. What it represents is a strange kind of social memory, one that is dictated by a strange kind of medium. The public-private issues for an archive of this sort reflect the unusual blurring of that line in our recent use of social media.

Historians and archivists collect and use a wide range of materials, from published official documents like court proceedings to ephemera. But, to return to Lincoln for the moment, many historical biographies draw on collections of public speeches, private letters, and sometimes diaries. It is worth noting here, that there is already some significant similarity between letters and diaries, the latter being, perhaps, a self-directed letter. One of the reasons we are able to rely on letters is that they have been widely distributed and individually archived—at least until collected at a later date. Many of these personal communiqués (that is, e-mails) are now archived for modern users of the web; in my case by Google. But the wide adoption of social media has encouraged a new transparency in these formerly private expressions.

It is still a window that is only partly open. Despite the hyperbole, only a relatively small proportion of the population regularly uses Twitter, for example. But this still represents a form of documentation of everyday communication that is novel and worthy of collection. Moreover, social media is still in its earliest stages; we will continue to want to share our lives with our social networks and larger publics, and online exchanges will continue to make up the bulk of those interactions. As remembrance agents and other forms of personal history systems come to wider use, we will have more to share, and more ways to do so[19]. As we increase our potential awareness of our surroundings, we become enmeshed in a co-surveillance network that changes the way we see the world, the way we see our past, and therefore who we are and can become.

Our memories are not our own. We always form our memories in tension with the views of those around us. They are negotiated until there is some common understanding, a narrative that we can agree upon. What happens when there is enough evidence to continually form and reform a personal history? Will the task of the future historian resemble that of the current mass marketer: making sense of millions of distributed conversations and forming an overall story from them? How we someday remember ourselves individually, and as a society, will be determined in large part by what we archive today, and the way we archive it.

[1] Kruegle, Herman, *CCTV Surveillance: Analog and Digital Practices and Technology.* 2 ed (Amsterdam: Butterworth-Heinemann, 2005), 298.
[2] Lasswell, Harold D., "The structure and function of communication in society," in L. Bryson (Ed.), *The communication of ideas: A series of addresses* (New York: Harper, 1948).
[3] "If you see something, say something," *http://www.mta.info/mta/security/index.html*; CNN's iReport, *http://www.ireport.com/* .
[4] Mann, Steve, Jason Nolan and Barry Wellman, "Sousveillance: Inventing and using wearable computing devices for data collection in surveillance environments," *Surveillance and Society* 1(3): 331-55.
[5] Brin, David. *The transparent society*. (New York: Basic Books, 1999).
[6] Schramm, Wilbur Lang, Steven H. Chaffee and Everett M. Rogers, *The beginnings of communication study in America* (Newbury Park: Sage, 1997), 36-37.

[7] Kirkpatrick, Marshall, “The future of social media monitoring,” *ReadWriteWeb* (April 15, 2009), *http://www.readwriteweb.com/archives/whats_next_in_social_media_monitoring.php*

[8] A study by the US Library of Congress suggested that audio CDs were susceptible degradation with age, and there have been suggestions that CD-Rs also had fairly short reliable lifetimes. See Shahani, Chandru J., Basil Manns and Michele Youket, “Longevity of CD media: Research at the Library of Congress,” Preservation Research and Testing Division, *http://www.loc.gov/preserv/studyofCDlongevity.pdf* (retrieved April 4, 2009).

[9] The difficulty of setting limits on the edges of “documents,” and the necessity of preserving the linked context, has long been known as a particular challenge for web archivists. See Lyman, Peter, “Archiving the World Wide Web,” in *Building a national strategy for digital preservation: Issues in digital media archiving* (Washington, DC: Council on Library and Information Resources and the Library of Congress, 2002), 41.

[10] The play on Warhol’s phrase has been attributed to a number of folks, and David Weinberger provides a link to a version on Usenet in 1997: Weinberger, David. “Famous to 15 people,” *Joho the Blog* (July 23, 2005), *http://www.hyperorg.com/blogger/mtarchive/004264.html*

[11] And it has been suggested that a similar system could extend the functionality of a proxy to help create personal, searchable archives: Rao, Herman Chung-Hwa, Yih-Farn Chen, Ming-Feng Chen, “A proxy-based personal web archiving service,” *ACM SIGOPS Operating Systems Review* 35(1), 2001, 61-72.

[12] Borges, Jorge Luis. “On exactitude in science,” in *Collected Fictions*, trans. Andrew Hurley. (New York: Viking, 1998), 325; Carroll, Lewis, *Sylvie and Bruno Concluded* (New York: Macmillan & Co., 1894), 169.

[13] Organizations are also retaining material more than they have in the past. Again, this is in part because they are able to—mail servers collect messages more effectively than carbon paper ever could have—but it is also a legal requirement for many large businesses. However, because of the organizational and legal barriers to accessing this material publicly, it is left aside in this discussion.

[14] Freedman, Michael J., Eric Freudenrthal, David Mazières, “Democratizing content publication with Coral,” in Proc. 1st USENIX/ACM Symposium on Networked Systems Design and Implementation *(NSDI '04)* San Francisco, CA, March, 2004.

[15] Schachter, Joshua, “On URL shorteners,” *Joshua Schachter’s Blog*, April 3, 2009, *http://joshua.schachter.org/2009/04/on-url-shorteners.html*; “TinyURL,” *ArchiveTeam* wiki, *http://archiveteam.org/index.php?title=TinyURL*, accessed April 5, 2009.

[16] Iyer, Sitram, Antony Rowstron, and Peter Druschel, “Squirrel: A Decentralized peer-to-peer web cache,” In Proc. 21st Annual Symposium on Principles of Distributed Computing, Monterey, 2002, 213-222.

[17] See Cooper, Brian F. and Hector Garcia-Molina, “Peer-to-peer data trading to preserve information,” *ACM Transactions on Information Systems* 20(2): 133-170.

[18] My dissertation suggests that some of the large group weblogs, and Slashdot in particular, brought obscure sites to the attention of a mass audience. Clearly Digg does this as well, as do popular blogs like Boing Boing. Halavais, Alexander, *The “Slashdot effect”: Analysis of a large-scale public conversation on the world wide web*. Unpublished dissertation, University of Washington, 2001.

[19] Rhodes, Bradley J., “The wearable remembrance agent: a system for augmented memory,” Proc. Of the 1st IEEE International Symposium on Wearable Computers, Cambridge, Massachusetts, 1997; Cowley, Paula, Jereme Haack, Rik Littlefield, and Ernest Hampson, “Glass box: capturing, archiving, and retrieving workstation activities,” Proc of the 3rd ACM workshop on Continuous Archival and Retrieval of Personal Experiences, Santa Barbara, California, 2006.