DOI: 10.1145/1592761.1592794

BY FRANK MCCOWN, CATHERINE C. MARSHALL, AND MICHAEL L. NELSON

Why Web Sites Are Lost (and How They're Sometimes Found)

THE WEB IS IN CONSTANT FLUX – new pages and Web sites appear daily, and old pages and sites disappear almost as quickly. One study estimates that about two percent of the Web disappears from its current location every week.[2] Although Web users have become accustomed to seeing the infamous "404 Not Found" page, they are more taken aback when they own, are responsible for, or have come to rely on the missing material.

Web archivists like those at the Internet Archive have responded to the Web's transience by archiving as much of it as possible, hoping to preserve snapshots of the Web for future generations.[3] Search engines have also responded by offering pages that have been cached as a result of the indexing process. These straightforward archiving and caching efforts have been used by the public in unintended ways: individuals and organizations have used them to restore their own lost Web sites.[5]

To automate recovering lost Web sites, we created a Web-repository crawler named Warrick that restores lost resources from the holdings of four Web repositories: Internet Archive, Google, Live Search (now Bing), and Yahoo;[6] we refer to these Web repositories collectively as the *Web Infrastructure* (WI). We call this after-loss recovery *Lazy Preservation* (see the sidebar for more information). Warrick can only recover what is accessible to the WI, namely the crawlable Web. There are numerous resources that cannot be found in the WI: password protected content, pages without incoming links or protected by the robots exclusion protocol, and content hidden behind Flash or JavaScript interfaces. Most importantly, WI crawlers do not have access to the server-side components (for example, scripts, configuration files, databases, among others) of a Web site.

Nevertheless, upon Warrick's public release in 2005, we received many inquiries about its usage and collected a handful of anecdotes about the Web sites individuals and organizations had lost and wanted to recover. Were these Web sites representative? What types of Web resources were people losing? Given the inherent limitations of the WI, were Warrick users recovering enough material to reconstruct the site? Were these losses changing their behavior, or was the availability of cached material reinforcing a "lazy" approach to preservation?

We constructed an online survey to explore these questions and conducted a set of in-depth interviews with survey respondents to clarify the results. Potential participants were solicited by us or the Internet Archive, or they found a link to the survey from the Warrick Web site. A total of 52 participants completed the survey regarding 55 lost Web sites, and seven of the participants allowed us to follow-up with telephone or instant messaging interviews. Par-

ticipants were divided into two groups:
1. *Personal loss:* Those who had lost (and tried to recover) a Web site that they had personally created, maintained or owned (34 participants who lost 37 Web sites).
2. *Third party:* Those who had recovered someone else's lost Web site (18 participants who recovered 18 Web sites).

What Was Lost (and Found)?

One might imagine that the lost Web sites occupy a minor niche, that they are small or have a very limited audience, and do not represent significant financial value. The survey results contradicted these expectations. Nor were the Web sites limited to simple static pages; they were often complex, with socially or programmatically generated content. Furthermore, the losses were extensive, usually involving entire sites. Recovery was equally complicated, owing not only to deep Web or Web 2.0 content, but also because there were sometimes gaps between when the Web site vanished and when the recovery commenced.

The lost Web sites covered a broad range of subjects (Table 1). Web sites about hobbies and interests ran the gamut from Frank Sinatra to Indian movies. Educational sites covered an array of subjects such as humanistic Judaism, women's health, and ancient Roman history. Many of the family/personal Web sites contained photos, articles or blog postings, and other content of emotional value. One participant described his lost content as "sort of my personal blog, so it is valuable to me for the same reason that old photos are valuable. For sort of nostalgia. Looking back and see-

Lazy Preservation and Warrick

Frank McCown and Michael L. Nelson

As the Web becomes a hub for our daily activities, curation and preservation of Web-based material imposes an increasing burden on individuals and institutions. Conventional Web preservation projects and techniques require a significant investment of time, money, and effort and thus are applicable only to collections of acknowledged value. The limited scope of such projects may leave many potentially important Web collections unprotected.

Lazy Preservation addresses the recovery of these unprotected collections.[3,4] Lazy Preservation does not require an institutional commitment to a particular archive; rather it is achieved by the ad hoc, distributed efforts of individual users, web administrators and commercial services. This strategy takes advantage of a growing *Web Infrastructure* (WI), which includes the harvested holdings of search engine companies (e.g., Google, Yahoo, Live Search), non-profit organizations (e.g., the Internet Archive's Wayback Machine) and large-scale academic projects (e.g., CiteSeer, NSDL). The WI refreshes and migrates web content in bulk as a side-effect of user services; these holdings can be mined as a useful, but passive preservation service. Although recovery results for a specific object sometimes can be disappointing, the aggregate performance for a complete website is usually very good. Like RAID (Redundant Arrays of Inexpensive Disks) systems, where reliable storage is built on top of individually unreliable disks, the WI provides a dependable resource for content recovery, even if individual elements of the resource are missing. However, unlike RAIDs, the WI elements are not under our control.

Warrick is a web-repository crawler which uses Lazy Preservation principles to recover lost websites.[3] Warrick "crawls the crawlers;" it begins with a seed URL of a lost website and makes requests to four web repositories: Internet Archive, Google, Live Search, and Yahoo. Of these repositories, only the Internet Archive retains the web resources in their original format; the other repositories may store modified versions of non-HTML content such as images, PDF, and Microsoft Office documents. The most recent version of the resource or the resource stored in its original format is saved to disk, and HTML resources are mined for links to other missing content. Warrick continues to recover resources until its queue is empty; checkpoints are set if daily query quotas are exceeded. A queuing system that runs Warrick jobs on a network of machines hosted at Old Dominion University[a] currently averages approximately 100 jobs a month.

Initial experiments with Warrick have confirmed the utility of using multiple web repositories to reconstruct websites.[3]

[a] http://warrick.cs.odu.edu/

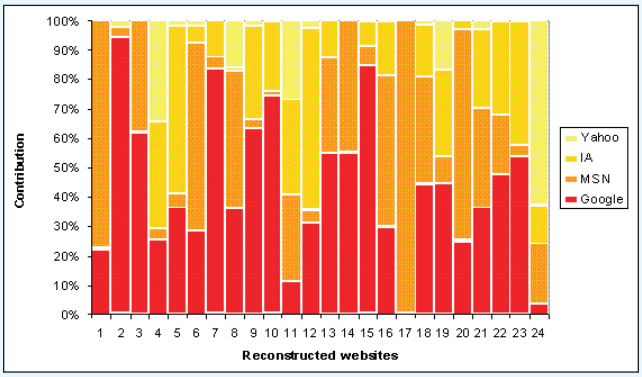


Figure 1. Web repositories contributing to each website reconstruction.

Figure 1 shows how the four web repositories contributed widely varying amounts to reconstructions of 24 websites. For example, Google's cache provided 95% of the resources for recovering website 2 but only 22% for website 1. More extensive experiments reconstructing 300 randomly selected websites over a period of three months have shown that on average 61% of a website's resources (77% textual, 42% images and 32% other) could be recovered if the website were lost and immediately reconstructed.[2]

One challenge of Lazy Preservation is that the WI only has access to the surface web; deep web content and website server components (CGI scripts, databases, etc.) cannot be recovered in the event of a loss. We are currently investigating methods for recovering the server components of a website by breaking the components into smaller encoded pieces (using erasure codes[5]), suitable for injecting into crawlable portions of the site. For example, a repository's source code could be encoded and stored in the HTML pages it produces. When the HTML pages housing the server components are discovered and stored by the WI, recovering a subset of the pages allows the entire set of server components to be recovered.

References

1. F. McCown, A. Benjelloun, and M. L. Nelson. Brass: A queueing manager for Warrick. In IWAW '07: Proceedings of the 7th International Web Archiving Workshop, June 2007.
2. F. McCown, N. Diawara, and M. L. Nelson. Factors affecting website reconstruction from the web infrastructure. In JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries, June 2007, 39-48.
3. F. McCown, J. A. Smith, M. L. Nelson, and J. Bollen. Lazy preservation: Reconstructing websites by crawling the crawlers. In WIDM '06: Proceedings from the 8th ACM International Workshop on Web Information and Data Management, 2006, 67-74.
4. M. L. Nelson, F. McCown, J. A. Smith, and M. Klein. Using the web infrastructure to preserve web pages. International Journal on Digital Libraries, 6(4), 2007, 327-349.
5. M. O. Rabin. Efficient dispersal of information for security, load balancing, and fault tolerance. Journal of the ACM, 36(2), 1989, 335-348.

Table 1. The nature of the 55 Web sites (categories are not exclusive)

Category	Personal loss *N* = 37	Third party *N* = 18	Examples
Hobby	16 (43%)	5 (28%)	Photographs of cemeteries
Family / personal	12 (32%)	1 (6%)	Political blog and article archive
Education / training	9 (24%)	8 (44%)	Typography and design
Commercial	8 (22%)	2 (11%)	Irrigation technology
Entertainment	7 (19%)	3 (17%)	Christian music e-zine
Professional	7 (19%)	1 (6%)	Painting business
Other	3 (8%)	1 (6%)	Opera commentary

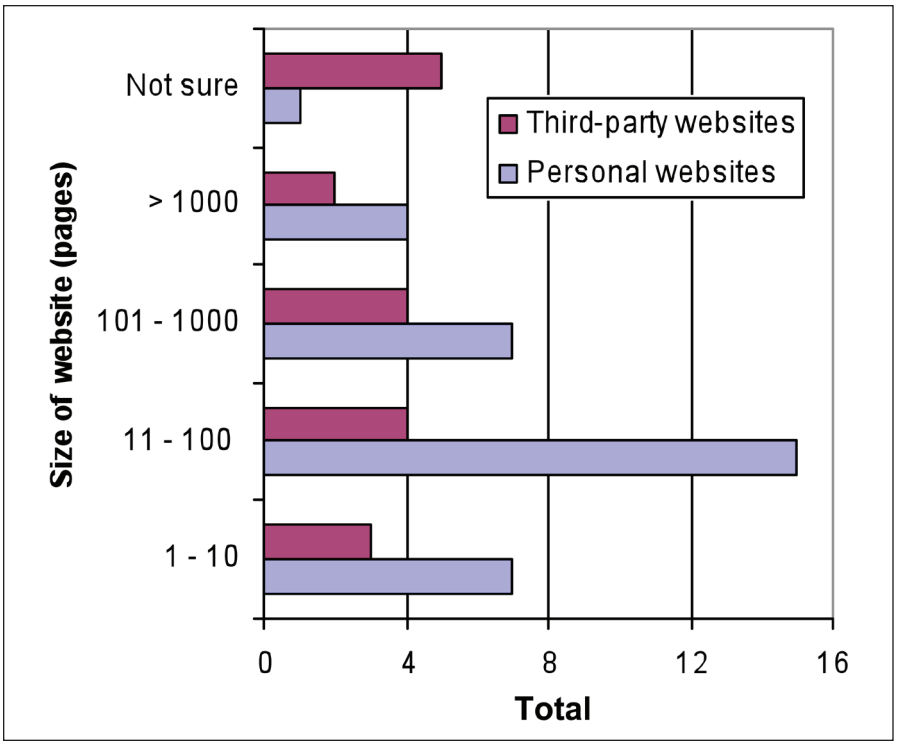
ing how you've grown. Reminiscing."

A surprising number of lost sites were of commercial value. Some were used directly to sell products or services, for example an e-commerce site for a major jewelry retailer. Others were geared towards marketing or communication. One Web site served as the primary information source and social nexus for a city-wide kickball league and another as the primary marketing tool for an irrigation business. Several Web sites respondents categorized as entertainment or professional were also of commercial value, and loss of the Web site meant loss of revenue in one form or another for the owner. The owner of a small house-painting business told us that his Web site "is on my business cards; it's on all my signs. And I've gotten people from Ohio... from Chicago [who] get my Web address, look at my jobs, and call me because they're coming out to buy a condo."

A majority of the Web site owners (67%) paid for hosting services. Four owners had their sites hosted on free services like geocities.com; three had a Web site on their university's Web server; one used an ISP, and one used his own personal server.

The size, makeup, and audience of the lost Web sites varied considerably. More than half were extensive resources: 29% percent had between 11 and 100 Web pages and 38% were larger than 100 Web pages (Figure 1). Furthermore, many of them had user-contributed or dynamic content. 21% percent of the sites had a blog, 6% had a wiki, and 31% had a forum. Although many of the Web sites were a collection of static files, 43% of them contained content that was generated using CGI, PHP, databases, or other programmatic means. The effect of the loss may have been widespread as well, extending far beyond the original owner: more than half of the participants (56%) believed the Web sites were used by at least 50 people before the loss.

Figure 1. Distribution of lost Web site sizes (number of Web pages).



The losses suffered were substantial. Over 90% percent of the participants claimed the Web site of interest was completely lost or almost completely lost. Yet despite the magnitude of loss and apparent value of the Web sites, the losses were not always discovered immediately. Although 65% of the participants discovered the loss in a week or less, 29% required at least a month to discover the loss. This temporal gap is a significant obstacle to recovery because inaccessible resources may begin to drop out of search engine caches just a few days after they are no longer accessible from the Web;[6] the window of opportunity to recover the lost resources may have passed for more than a quarter of the participants.

The problem was even worse for those involved in third-party recoveries; 65% of those who recovered someone else's Web site did not learn of the Web site loss until more than a month had gone by. It was not always clear to these respondents that the loss was not due to a temporary outage: "They thought [the site outage] was because of their Web host company... Then the staff changed over and it just became this line of, um, I guess not keeping a track record of what's going on."

Once a loss was discovered and indeed perceived as such, were respondents able to recover the portions of the Web site that mattered to them? Thirty-three of the 52 participants had finished trying to recover their lost site or someone else's lost site before they took our survey. Of these, almost half were able to recover most or nearly all of the lost site (Figure 2). Unfortunately, 52% of the participants said there was an "important" part of their Web site which could not be recovered. Half of the respondents indicated the items permanently lost were the server-side components of their sites; others claimed their mp3s, forums, images,

and other content were unrecoverable. According to one participant, "[There were] lots of missing holes in the content which is very frustrating. Archive.org didn't catch everything." Another participant noted that there was no way to tell whether he had recovered all of his blog posts: "There're literally hundreds of posts. And not to mention the fact that I wouldn't even necessarily have a perfect memory of whether a post existed or not."

The Blame Game

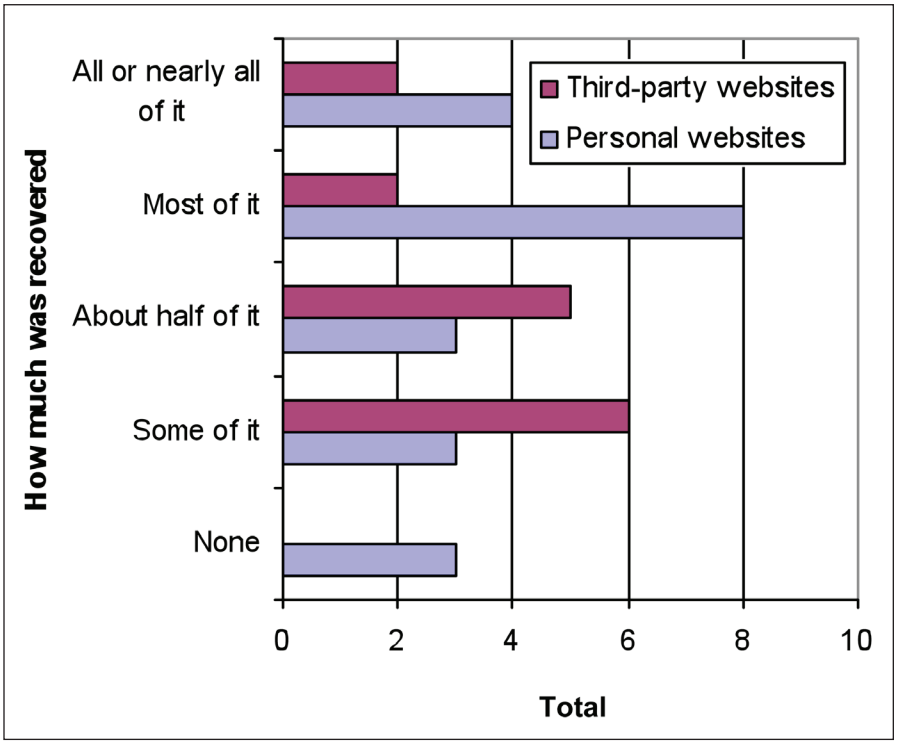
It's easy to see that multiple parties may be involved in a Web site's disappearance: the owner is not necessarily the designer, nor does the Information Architect have any control over the ISP's policies. Larger institutional, social, and legal issues may come into play as well.

Accordingly, the participants' reactions to losing their sites were mixed. One third of the participants made it clear the loss was major: "Devastated. Months and months of code was lost." Others complained of important content that was gone, loss of countless hours of work, and interruption of "very important sales." The other two thirds thought the loss was less severe or minimal; one participant said that although the lost site only affected himself, he "felt kind of sad" since it was the very first Web site he had ever created. A few seemed more ambivalent, sardonically shrugging off the loss: "I'm sure my future biographers [will] lament the loss."

When asked why their Web sites were lost, 43% of the participants blamed the hosting company. Free hosting companies deleted two of the Web sites for unknown reasons; six more sites were lost when the hosts experienced disk failure and had a corrupted backup or no backup at all. Several hosting companies were apparently the victims of hackers or viruses, and several others went out of business and removed their customers' content without notice. One ISP was hosting a Web site dealing in pirated software and movies, and the respondent's site was lost along with the offending site as the result of a Swedish police raid.

Other sites were lost through owner negligence. One Web site was deleted months after the owner forgot to renew his hosting services (the renewal notification was inadvertently caught by his spam filter). Another owner accidentally deleted her Web site. A few others experienced hardware failures when hosting the sites on their own servers. In one case, the owner of the site purposefully let the site die out but then changed his mind several years later.

Figure 2. Distribution of perceived recovery.



Sites may also be lost through changing circumstances or relationships. When institutional affiliations change, Web sites may get lost in the shuffle; one site owner forgot to move the site to another location when he left school and the system administrators deleted his account. Another site was lost when the site's owner and site's maintainer had a falling-out: "I contacted [my friend who had developed the site] and he said that if I gave him a hundred dollars an hour that he could go ahead and pull it up for me and get it back online. And I thought that was kind of a slap in the face." Two Web sites were recovered by interested third parties when the sites' owners died and left no backups. Two other sites were lost when the companies they represented went bankrupt, one the victim of the dot-com bubble. Finally, sometimes larger social forces are at work: a site documenting the medicinal and recreational cultivation of marijuana was taken down by Canadian police; the recovered site was never re-hosted but instead used by the recoverer as a personal resource.

Backups, or Lack Thereof

Our survey revealed that many individuals did not backup their Web sites and relied instead on the hosting company to protect their files from loss. Almost 60% of the participants never created a single backup of their Web sites, and of the 11 individuals that did, a third of them performed the backup process manually (which is error-prone). Most found their backups somewhat (73%) or very (18%) useful in recovering their Web site, and in these cases Warrick was able to supplement the recovery by finding additional lost resources.

Participants who paid for hosting services tended to have higher expectations of their hosting provider than those who received free services. One participant lashed out at the hosting company's "incompetent system admins", and another voiced his frustration that the hosting company never replied to any of his emails.

Although most individuals know that they should backup their data, they rarely do. It is not uncommon

for individuals, even those who work on storage backup techniques, to admit they do not backup their personal files.[1] Although researchers have proposed a number of methods to make backup simple and affordable for the masses,[1] such systems are not yet in widespread use. Commercial backup systems are prohibitively expensive for some (Backup.com offers 1GB of storage for $15 a month), and so backup is therefore generally confined to the organization, not the individual. One of our respondents who did not back up his Web site, even though it was hosted on his own server, exclaimed, "Whose fault is it? I mean, is it the user's fault for not backing up? Or is it technology's fault for not being more tolerant and failsafe, right? In ten years, maybe hard drives and PCs will be so invincible and the Internet will be so pervasive that the concept of backing up will be quaint, right?"

When they do create backups, individuals tend to backup their important files using a number of ad hoc techniques (such as, emailing themselves files, retaining old systems with important files, or spreading the content across free services to mitigate risk) which may or may not allow complete recovery in the face of disaster.[4,5] Because it is so rare for a hard drive to crash or for a Web hosting company to go out of business, individuals are not sufficiently motivated to keep their important files backed-up. For those performing third-party reconstructions, the owners' backup practices are inconsequential since third parties do not normally have access to private backups.

Doing Things Differently

Given the nature of some of their losses, we might expect respondents to be quick to assert that they are going to change their ways. Indeed, several participants said they were transferring their Web sites to hosts that promised reliable backups. Others said they would continue to use free hosting services, but only services from larger companies with the expectation that the larger companies will be more responsible. Several participants said they would perform backups more regularly, use automated backup tools, or keep more backup copies, even when using another Web hosting company. One participant who lost the server components of his dynamic Web site said he was going to backup both the server files and perform a full crawl of the Web site, just in case the server files would not run in the future. In spite of these good intentions, several respondents had not yet implemented their new failsafe strategies in the four months between the survey and interviews.

Other participants, however, expressed they would not do anything differently to protect their Web sites. The participant who deleted his Web site said he would just be "a tad more careful with regard to which directory [he was] in." Another said he was going to do backups "sometimes" as opposed to never. One participant who lost a portion of a large community site when the server crashed said there was not much he could do differently since he used an automated backup before the loss.

Conclusion

Given the diversity of our respondents' Web sites and their motivations for using the WI to restore them, we can surmise that trends that are common among them represent general characteristics of digital loss. Four important findings are:

1. The `long tail' effect is demonstrated by the Web sites and respondents' motivations for restoring them. Individuals are restoring deep resources that pertain to relatively narrow domains, be they personal, topical, or commercial; these sometimes-esoteric resources are adjudged to be of sufficient value to warrant the restoration effort.

2. People place themselves at considerable risk for loss, partly through circular reasoning (the fallacy of the safe local copy), partly through lack of familiarity with service provider policies and practices, and partly through normal kinds of benign neglect carried over from caring for physical materials (for example, the photos in the cardboard box under the bed).

3. Web site salvage that relies on current WI may become more unreliable as we move toward Web 2.0, where content is dynamic, socially generated, or inaccessible to crawlers.

4. Finally, as we create more and more digital content as a normal part of our everyday activities, it seems that we will have less time to curate what we have already, not more. Furthermore, our expectations of automatic data safety will increase. If we don't backup our files now, we shouldn't expect to do so in the future.

The survey results having important implications for personal digital preservation, for the WI, and for Lazy Preservation tools like Warrick. As the Web becomes more capable and complex, and as we begin to live a greater portion of our lives online, both the WI and the means to extract content from it will have to become more inclusive. Technology to assist people in the onerous task of preserving the digital materials that comprise quotidian (yet undeniably important) human activities must interleave seamlessly with these activities; people who don't find time to backup their Web sites are not apt to adopt anything that requires extra thought and planning. The payoff for curation (the ability to look at digital photos in fifty years) is too far downstream to make anything other than benign neglect seem worthwhile. Tools like Warrick (after-loss recovery) may align more closely with human nature than preservation applications that require up-front effort. C

References

1. Cox, L. P., Murray, C. D., and Noble, B. D. Pastiche: Making backup cheap and easy. *SIGOPS Operating Systems Review 36*, SI, (2002), 285-298.
2. Fetterly, D., Manasse, M., Najork, M., and Wiener, J. A large-scale study of the evolution of Web pages. In *Proceedings of WWW '03*, (2003), 669-678.
3. Kahle, B. Preserving the Internet. *Scientific American*, (Mar. 1997), 82-83.
4. Marshall, C., Bly, S., and Brun-Cottan, F. The long term fate of our personal digital belongings: Toward a service model for personal archives. In *Proceedings of IS&T Archiving 2006*, (2006), 25-30.
5. Marshall, C., McCown, F., and Nelson, M. L. Evaluating personal archiving strategies for Internet-based information. In *Proceedings of IS&T Archiving 2007*, (2007), 151-156.
6. McCown, F., Smith, J. A., Nelson, M. L., and Bollen, J. Lazy preservation: Reconstructing Websites by crawling the crawlers. In *Proceedings of ACM WIDM '06*, (2006), 67-74.

Frank McCown (fmccown@harding.edu) is an assistant professor of computer science at Harding University, Searcy, AR.

Catherine C. Marshall (cathymar@microsoft.com) is a senior researcher at Microsoft Research, Silicon Valley.

Michael L. Nelson (mln@cs.odu.edu) is an associate professor of computer science at Old Dominion University. Prior to joining ODU, he worked at NASA Langley Research Center.