Society | DOI:10.1145/2843553

Esther Shein

Preserving the Internet

Is the Internet ephemeral by its nature, or can it be archived?

THERE WAS A time when an interested party could find five White House press releases online detailing the number of countries participating in the "Coalition of the Willing" in support of the 2003 invasion of Iraq. That list drew controversy, however, because when the war did not go so well, countries like Togo and Costa Rica objected to having their names associated with it. Consequently, the data was manipulated, says Kaylev Leetaru, who co-authored the report "Airbrushing History, American Style."

Today, two of those five documents are unchanged, two have been deleted, and one is accessible but has been edited from its original form, according to Leetaru. "There was a lot of manipulation, and what came out of that was the White House was altering press releases on a regular basis,'' he says. "It came out that they didn't view these as government documents, but rather propaganda material, so there was no reason to mark that a change was made."

Leetaru, creator of the GDELT (global database of events, language and tone) project, which monitors broadcast, print, and Web news worldwide in over 100 languages, also found an instance in which the White House issued a statement from then-President George Bush about the economy improving. "When it didn't, they went back and edited a press release from a few years prior so they could say, 'look, we were right.'" Altering history online is often done by the government and elected officials when that information turns out to be factually incorrect, Leetaru says.

"For government and elected government officials in particular, they don't want things preserved because people can go back and see what the government actually said, so there's actually a perverse effort not to preserve it,'' says Leetaru, who is also a senior fellow at George Washington University. Such selective archiving is not limited to the government, however; many companies also do not archive information "because there is no benefit and it would create confusion to have old material up,'' he says.

Additional storage awaiting use by the Internet Archive.

IMAGE BY JASON "TEXTFILES" SCOTT

Not to mention embarrassment. Leetaru also cites the (in)famous statement "there is no market for home computers," which can be found online in the 1988 *University of Michigan Computing News*, Volume 3, (http://bit.ly/1PZ7k74).

It is not surprising some people might not want certain statements preserved for posterity, and that websites are updated and information altered; after all, the Internet is ephemeral by nature. Yet in doing so, we are doing a disservice to future generations by not passing on relevant societal information that provides an accurate view of the past. Past ACM president Vint Cerf, long considered the "father of the Internet" and now a vice president and chief Internet evangelist at Google, expressed concern last year that we are in a "digital Dark Age," and that future generations will have little or no record of the 21st century, because so much data is kept in a digital format and technology advances so quickly that old files will be inaccessible.

Leetaru notes that in the European Union there is a "Right to be Forgotten" law that basically says if someone wants information on them removed from the Internet, major search engines like Google must delete it from their search index. Leetaru worries other countries might adopt similar regulations. If something cannot be found on Google, he points out, it essentially does not exist to Internet users, and history is being rewritten. "So if a website is removed from Google and all the major search engines, for all intents and purposes it's removed from the world. That's a really scary situation in terms of our collective remembrance of society."

However, "The Web was never designed for being archived,'' observes Brewster Kahl, founder and digital librarian of the Internet Archive, a not-for-profit digital library whose goal is to preserve the Internet's past for the use of future historians. The average lifespan of a Web page is a little under 100 days before it is changed and/or deleted, he says.

Kahl felt strongly that as the Web grew, the Internet Archive should help provide universal access to all knowledge. "Libraries were useful in the past, and as we're changing a publishing system from paper-based to digital, we should do the same,'' says Kahl. The idea was further fleshed out to provide access to materials no longer available online "and to make it so people could compute on these materials."

About 140 people work at Internet Archive scanning hundreds of books per day, worldwide. The organization receives about $12 million per year in funding, mostly from libraries paying the Internet Archive for its services, as well as donations from foundations, he says.

Kahl also created the Wayback Machine, whose name is a nod to a fictional time machine from the "Rocky and Bullwinkle" cartoon show in the 1960s. It is a three-dimensional index that archives and allows browsing of current and older Web documents. "We started by being a kind of robot that crawled everything we could find," says Kahl, "but it's evolved since then." Some 1,000 librarians build custom Web collections from about 350 different libraries, museums, and archives worldwide, "so now there are lots of people helping select and make sure the Wayback Machine has the right things in it,'' he says. "The Web is not just one thing anymore; it's lots of different collections. At least, that's the way we think of it."

Today, 600,000 people access the Wayback Machine each day, searching for approximately 450 billion Web objects, which include images. About one billion pages are added each week, Kahl says. Leetaru has collaborated with the Internet Archive and Flickr to extract images from 600 million digitized book pages that date back 500 years from over 1,000 libraries worldwide and make them all browseable and searchable via both the metadata of the original book and the text surrounding each image.

Among the other organizations working with the Internet Archive is the U.S. Library of Congress (LOC), which took some heat early last year for failing to digitize its collections. Since 2000, the LOC has been preserving Web content not only from its own Website, but also the sites of other organizations and citizens as part of its Web Archiving program (http://www.loc.gov/webarchiving/), says Abbie Grotke, lead information technology specialist with the LOC's Web Archiving Team. In addition to event-based archives on topics including the U.S. national elections, the Iraq War, and the events of September 11, 2001, the team's focus is on preserving content deemed "at risk." Grotke says this typically includes an event that has occurred but the documentation of it is no longer needed, such as campaign websites that are taken down after an election. It might also include government documents from unstable regions around the world.

The LOC archives different types of collections, such as online news sources, Web comics, and folklore, which are selected by "recommending officers," Grotke says. The team uses a custom

The Wayback Machine is a three-dimensional index that archives and allows browsing of current and older Web documents.

Milestones

Yelick Receives ACM/IEEE Award

Katherine Yelick, a professor of Electrical Engineering and Computer Sciences at the University of California at Berkeley and faculty scientist and Associate Laboratory Director for Computing Sciences at Lawrence Berkeley National Laboratory, was awarded the 2015 ACM/IEEE Computer Society Ken Kennedy Award for innovative research contributions to parallel computing environments that have been used in both the research community and in production environments.

Yelick, who has authored more than 170 technical papers and reports on parallel languages, algorithms, libraries, architecture, and storage, also was cited for her strategic leadership of national research libraries, and for developing novel educational and mentoring tools.

Yelick's work has improved the programmability of high-performance computing through innovations to parallel languages and runtime systems. Her contributions to compiler research and open source software were key to the success of partitioned global address space. She also developed automatic performance tuning techniques and runtime systems that maximize performance across a variety of computer architectures.

An ACM Fellow, Yelick was named 2013 Athena Lecturer by the ACM Women's Council (ACM-W). A member of the National Academies of Sciences, Engineering, and Medicine's Computer Science and Telecommunications Board, Yelick has served on the California Council on Science and Technology and on the University of California Science and Technology Council.

ACM and IEEE co-sponsor the Kennedy Award to recognize contributions to programmability and productivity in computing and significant service or mentoring contributions.

curator tool it built called Digiboard to nominate URLs to a collection, manage them, and track project activities. The tool is also used to request permission to crawl a particular site.

Universities are also making efforts to archive their websites. The Massachusetts Institute of Technology (MIT) has just started doing so, since it did not previously have the tools and technology in place, says Kari Smith, digital archivist at the MIT Libraries' Institute Archives and Special Collections, which uses an international file standard called WARC (Web ARChive) to combine different digital resources into an archival file.

MIT has eight full-time archivists working to preserve not only internal Websites, but also reports and attachments that might come from an administrative or faculty office, Smith says. "We know every semester information changes, because information on a website is about what's happening in a particular semester, so we look at dates that are most likely to change,'' such as those for specific courses. Although some sites change daily, "from an archival perspective, the question is, what can we capture on a daily basis and what is the future value of that for historical purposes?"

Professional archivists at the university decide what gets preserved, she says. "It's a collaborative decision based on our resources and what staffing and technology do we have available."

Other organizations also have as their mission the preserving information for future generations, among them the International Internet Preservation Consortium (IIPC) (http://www.netpreserve.org/) and Perma.cc (https://perma.cc), which helps create permanent links to online sources cited in works by scholars, courts, and journals.

The Digital Public Library of America (DPLA) works with Perma.cc and the Internet Archive, and also preserves the metadata it brings into its online repository through exhibitions, says Kenny Whitebloom, manager of special projects. "We have a small array of digital objects that we've acquired permissions to reproduce online,'' he explains. "We further facilitate access to those materials and help in their preservation on our site in that way, but we don't have a mandate to preserve Web pages or links."

> **"The World Wide Web is effectively infinite and has parts and corners that can't be archived, and there's large swaths of it that shouldn't be archived, that weren't meant for the ages."**

Nearly all Web archivists say it is virtually impossible to preserve all the content on the Web.

"The Web is so massive you can't crawl and digest everything," says Leetaru, "but I'm creating a very targeted collection of worldwide news media," especially local media, and providing URLs to make a permanent copy.

Leetaru does not hesitate when asked if enough government organizations and companies that should be preserving content for future generations are doing their due diligence. "Absolutely not … the average company doesn't care at all. Most companies delete content when it gets old. If you make a product you don't want to keep marketing material [online] because it will create confusion, so they delete material as part of their routine processes." Whereas it used to be costly to put that information on the Web, now companies just want to "clean up" their sites, he says.

Archivists generally do not preserve everything, says Smith, and resources are limited. As a result, the priority is on information that is ephemeral and the "output of people in the world that we can acquire and keep to tell the stories that are incredibly important—especially the stories that most often don't get told. We definitely want to make sure we're collecting actively about minority organizations or small groups that are having a big impact in their community," but may not get wide distribution.

Grotke also says it is possible to preserve portions of the Web, but notes that a lot of sites are constructed in a way that crawlers are unable to access them. "One reason is because of passwords, and current crawler technology is probably five to 10 years behind what the current Web is doing, so a lot of dynamic or flash-based content" such as art Websites that museums are trying to capture, is too complex and challenging for current tools, she says.

Kahl says the key is to capture what we want people in the future to know about the past.

"The World Wide Web is effectively infinite and has parts and corners that can't be archived, and there's large swaths of it that shouldn't be archived, that weren't meant for the ages,'' he maintains. "It's important to be selective. But on the other hand, we try to cast a very wide net, because we're interested in trying to help preserve the creative works of humankind."

Further Reading

Althaus, S. and Leetaru, K.,
University of Illinois Urbana-Champaign, 2008. Airbrushing History, American Style. http://courseweb.lis.illinois.edu/~katewill/spring2011-502/502%20and%20other%20readings/althous%20and%20leetaru%20Airbrushing%20History.pdf

"Blue Magic": A Review
***University of Michigan Computing News*, Volume 3, Number 19, p. 14, 1988, http://bit.ly/1NoZst2**

Venkataraman, B.,
The Race to Preserve Disappearing Data, *The Boston Globe*, May 17, 2015, http://bit.ly/1XGpKiS

O'Keefe, E.,
"Obama wants better digital archive of federal records," *The Washington Post*, Nov. 28, 2011, http://wapo.st/1LVlx45

Ghosh, P.,
"Google's Vint Cert warns of 'digital Dark Age,'" *BBC News*, http://bbc.in/1LVlB41

Executive Office of the President, "Memorandum for the Heads of Executive Departments and Agencies and Independent Agencies on Managing Government Records Directive," 2012, https://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf

Esther Shein is a freelance technology and business writer based in the Boston area.